# LEHD Infrastructure S2014 files in the FSRDC

by

**Lars Vilhuber**
**Cornell University and U.S. Census Bureau**

This paper is a revised version of *LEHD Infrastructure S2014 files in the FSRDC (CES 18-27) from May, 2018. A copy of the original paper is available upon request.*

**CES 18-27R September, 2018**

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact Christopher Goetz, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K028B, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov.

# Abstract

The Longitudinal Employer-Household Dynamics (LEHD) Program at the U.S. Census Bureau, with the support of several national research agencies, maintains a set of infrastructure files using administrative data provided by state agencies, enhanced with information from other administrative data sources, demographic and economic (business) surveys and censuses. The LEHD Infrastructure Files provide a detailed and comprehensive picture of workers, employers, and their interaction in the U.S. economy. This document describes the structure and content of the 2014 Snapshot of the LEHD Infrastructure files as they are made available in the Census Bureau's secure and restricted-access Research Data Center network. The document attempts to provide a comprehensive description of all researcher-accessible files, of their creation, and of any modifications made to the files to facilitate researcher access.

[*] This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It is released to inform interested parties of ongoing research and to encourage discussion of work in progress. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which was partially supported by the following National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant 5 R01 AG018854-02, and grants from the Alfred P. Sloan Foundation. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. The U.S. Census Bureau supports external researchers' use of these data through Research Data Centers (see www.census.gov/ces).

# Contents

# Chapter 1.
# Overview of LEHD Infrastructure

The Longitudinal Employer-Household Dynamics (LEHD) Infrastructure files available in the Research Data Center (RDC) are structured as individual components. A big-picture overview of it can be found in Abowd et al. (2006), which was published as Abowd et al. (2009). Figure 1.1 provides an overview of the flow of data elements through the LEHD data creation process.

Currently, the core outputs of the data creation process at LEHD are the Quarterly Workforce Indicators (QWI), shown in Figure 1.1, and the LEHD Origin-Destination Employment Statistics (LODES) data (also historically known as OnTheMap (OTM) data). The LEHD Infrastructure files in the RDC environment do not contain any information related to the disclosure limitation measures used at LEHD (for more information on the disclosure limitation techniques, see Abowd et al. (2006), Abowd, Stephens, and Vilhuber (2006), Haney et al. (2017), and Machanavajjhala et al. (2008) for a discussion).

After pulling the files from LEHD production archives, several research-related improvements are made to the files, addressing data access permissions issues (splitting or removing data), fixing minor data inconsistencies or updating documentation. Attempts are made to keep file structure and variable names consistent across Snapshot releases, though this is not always possible.

## 1.1 AVAILABILITY OF DATA

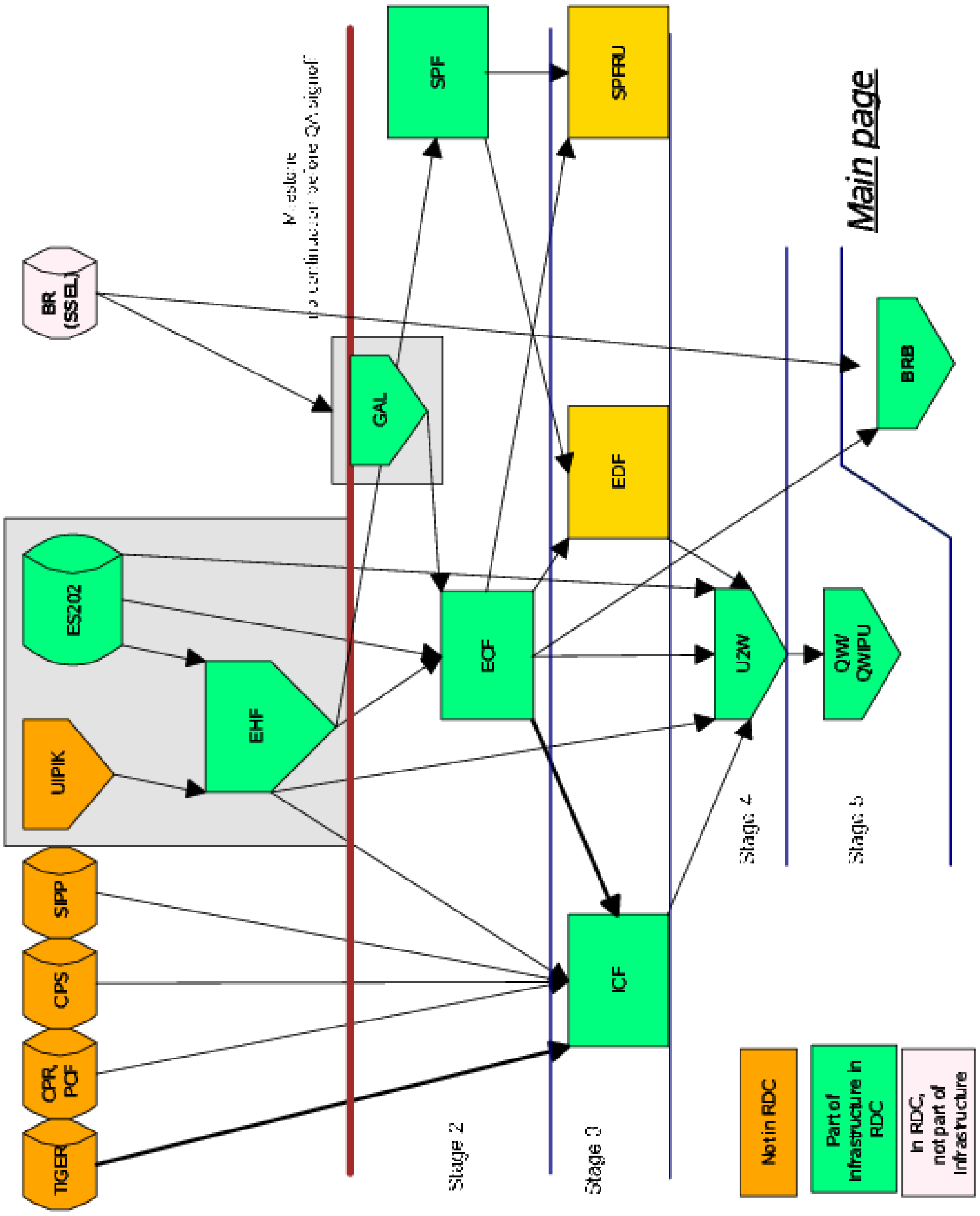
Availability of LEHD Infrastructure files is conditional on (i) the data files having been processed in the LEHD Production system, and subsequently integrated into the LEHD Infrastructure and (ii) permission for use in research having been granted by LEHD's state partner. The latter may vary over time, researchers should contact their FSRDC administrator for the most up-to-date information.

The standard Memorandum of Understanding (MOU) between the Census Bureau and its state partners precludes access to person and firm names and physical addresses as provided in the ES-202 data. As described below, there are geographic identifiers that are derived in the Geocoded Address List (GAL) that can be used for analysis and integrating data for appropriate and approved purposes. In addition to data provided by the states, and processed through the LEHD Production system, data provided by Office of Personnel Management (OPM) are also available (in experimental mode).

For the S2014Snapshot, all states (including the District of Columbia) had at some point been processed for the complete set of LEHD data files and integrated.[1] In general, LEHD Infrastructure files are available from 2000 onwards. However, the availability of historical data prior to 2000 varies significantly across states. Table 1.1 tabulates the availability data source (state UI or OPM) in the S2014snapshot (Figure 1.2 graphically depicts availability for UI/EHF data). Note that for certain states, availability of unemployment insurance (UI) files (as captured by the Employment History Files (EHF)) differs from historical availability of Quarterly Census of Employment and Wages (QCEW) files (as captured by the Employer Characteristics File (ECF)). Finally, a shorter time-series for the QWI indicates certain serious data issues interrupting the data series, sufficient to block publication of the official QWI, but

[1] States may and have dropped out of the voluntary state-federal partnership on which the LEHD Infrastructure draws.

Figure 1.1: Data flow view of LEHD Infrastructure



possibly without consequences for certain research uses. Data sources not currently available for the entire time period may become available in the next update to the LEHD Infrastructure, or as a revision to the current snapshot.

Table 1.1: Availability by data source

| | Start of data series | | |
|---|---|---|---|
| Data source | EHF | ECF | QWI |
| OPM | 2000Q1 | 2000Q1 | 2000Q1 |
| Alaska | 1990Q1 | 1990Q1 | 2000Q1 |
| Alabama | 2001Q1 | 2001Q1 | 2001Q1 |
| Arkansas | 2002Q3 | 2002Q3 | 2002Q3 |
| Arizona | 1992Q1 | 1992Q1 | 2004Q1 |
| California | 1991Q3 | 1991Q1 | 1991Q3 |
| Colorado | 1990Q1 | 1990Q1 | 1993Q2 |
| Connecticut | 1996Q1 | 1996Q1 | 1996Q1 |
| District of Columbia | 2002Q2 | 2000Q4 | 2005Q2 |
| Delaware | 1998Q3 | 1997Q1 | 1998Q3 |
| Florida | 1992Q4 | 1989Q1 | 1992Q4 |
| Georgia | 1994Q1 | 1994Q1 | 1998Q1 |
| Hawaii | 1995Q4 | 1995Q4 | 1995Q4 |
| Iowa | 1998Q4 | 1990Q1 | 1998Q4 |
| Idaho | 1990Q1 | 1990Q1 | 1991Q1 |
| Illinois | 1990Q1 | 1990Q1 | 1990Q1 |
| Indiana | 1990Q1 | 1990Q1 | 1998Q1 |
| Kansas | 1990Q1 | 1990Q1 | 1993Q1 |
| Kentucky | 1996Q4 | 1996Q4 | 2001Q1 |
| Louisiana | 1990Q1 | 1990Q1 | 1995Q1 |
| Maryland | 1985Q2 | 1985Q2 | 1990Q1 |
| Maine | 1996Q1 | 1996Q1 | 1996Q2 |
| Michigan | 1998Q1 | 1998Q1 | 2000Q3 |
| Minnesota | 1994Q3 | 1994Q3 | 1994Q3 |
| Missouri | 1990Q1 | 1990Q1 | 1995Q1 |
| Mississippi | 2003Q3 | 2003Q3 | 2003Q3 |
| Montana | 1993Q1 | 1993Q1 | 1993Q1 |
| North Carolina | 1991Q1 | 1990Q1 | 1992Q4 |
| North Dakota | 1998Q1 | 1998Q1 | 1998Q1 |
| Nebraska | 1999Q1 | 1999Q1 | 1999Q1 |
| New Hampshire | 2003Q1 | 2003Q1 | 2003Q1 |

(continued on next page)

Table 1.1 – Continued

| | Start of data series | | |
|---|---|---|---|
| Data source | EHF | ECF | QWI |
| New Jersey | 1996Q1 | 1995Q1 | 1996Q1 |
| New Mexico | 1995Q3 | 1990Q1 | 1995Q3 |
| Nevada | 1998Q1 | 1998Q1 | 1998Q1 |
| New York | 1995Q1 | 1990Q1 | 2000Q1 |
| Ohio | 2000Q1 | 2000Q1 | 2000Q1 |
| Oklahoma | 2000Q1 | 1999Q1 | 2000Q1 |
| Oregon | 1991Q1 | 1990Q1 | 1991Q1 |
| Pennsylvania | 1991Q1 | 1991Q1 | 1997Q1 |
| Rhode Island | 1995Q1 | 1990Q1 | 1995Q1 |
| South Carolina | 1998Q1 | 1998Q1 | 1998Q1 |
| South Dakota | 1994Q1 | 1994Q1 | 1998Q1 |
| Tennessee | 1998Q1 | 1998Q1 | 1998Q1 |
| Texas | 1995Q1 | 1990Q1 | 1995Q1 |
| Utah | 1999Q1 | 1990Q1 | 1999Q3 |
| Virginia | 1998Q1 | 1995Q3 | 1998Q1 |
| Vermont | 2000Q1 | 2000Q1 | 2000Q1 |
| Washington | 1990Q1 | 1990Q1 | 1990Q1 |
| Wisconsin | 1990Q1 | 1990Q1 | 1990Q1 |
| West Virginia | 1997Q1 | 1990Q1 | 1997Q1 |
| Wyoming | 1992Q1 | 1992Q1 | 2001Q1 |

The data underlying this table is attached to this document as CSV.

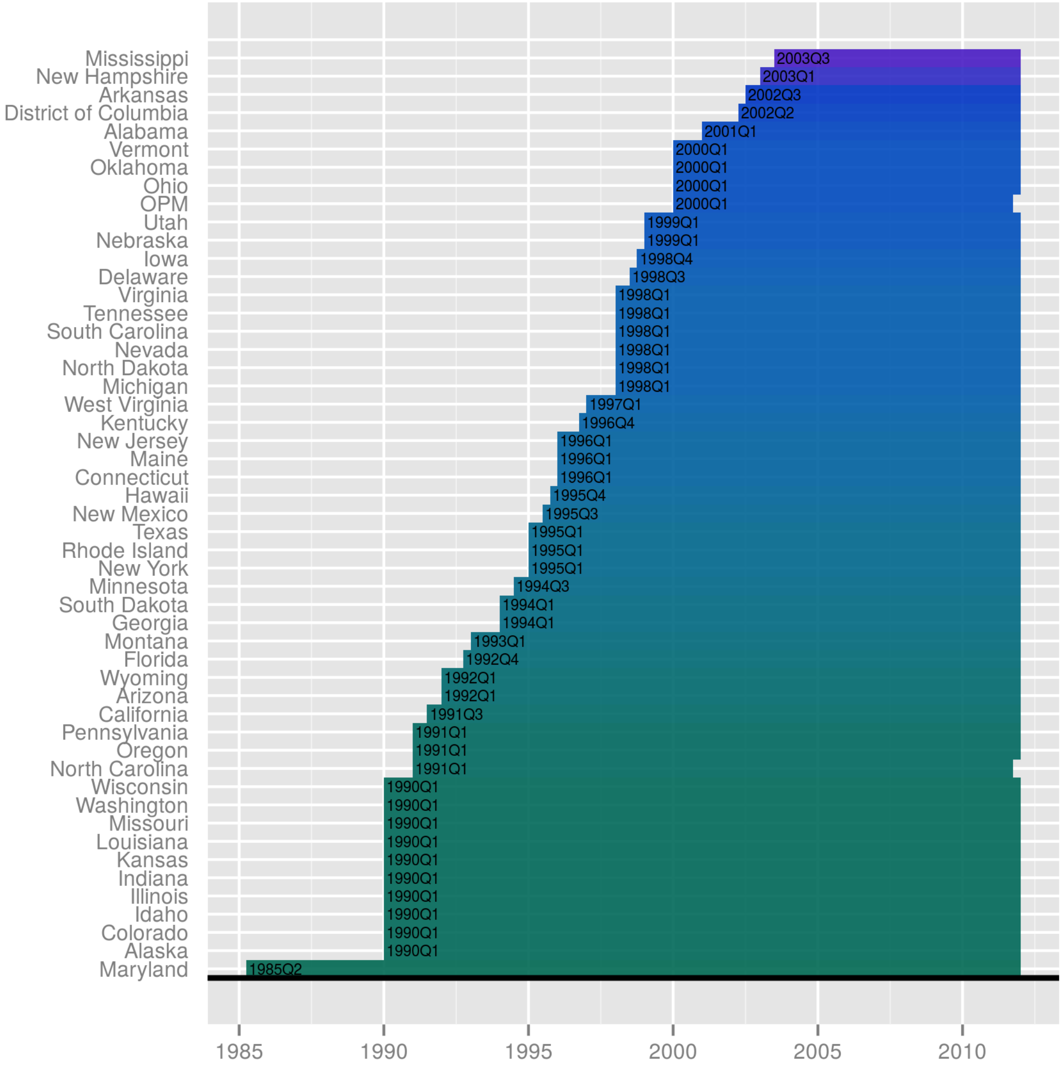
## 1.2 DISCLOSURE LIMITATION

Special disclosure and data use rules apply to analyses based on the micro-data from the LEHD Infrastructure file system. These data underlie the QWI and OTM, and research results are therefore subject to restrictions that ensure the LEHD disclosure limitation mechanism is not compromised. Model-based output is normally allowed. The chief disclosure officer for the RDC network will coordinate the reviews.

### Noise infusion

Disclosure limitation for the workplace-based statistics (QWI and LODES) uses noise infusion of the micro-data. The Disclosure Review Board (DRB) does not allow the release of any tabulations for sub-state geography that do not use the QWI noise infusion process. In addition, the required noise factors have not been placed on the RDC snapshot files as part of the DRB's normal rules limiting access to the specific parameters of its approved disclosure limitation methods.

Figure 1.2: Data availability (UI/EHF) by data source



### Tabular output

No tabular output from the LEHD infrastructure file system is allowed, unless explicitly authorized by the DRB. National or multi-state tables may be approved provided they do not compromise the protection system.

### Single state, multistate, and substate analyses

The underlying micro-data in the LEHD infrastructure file system were provided to the Census Bureau by states' Labor Market Information (LMI) offices under Memorandum of Understanding (MOU) negotiated with each state. This process is part of the Local Employment Dynamics (LED) federal/state partnership, and places additional restrictions on the results that may be published. Current members of the LED partnership are shown on the LEHD main web page. **??** discusses the implications for the project approval process. This section discusses the impact on the release of results.

- For all states, publicly disclosing a single state's data, or any sub-state information such as Metropolitan Statistical Area (MSA) or Core-Based Statistical Area (CBSA), in identifiable form requires the permission of the state's LMI agency. Each identified state shall be allowed to **review and approve release** prior to publication or distribution beyond authorized persons (i.e., "the public"). Note that in **contrast to the usual Census Bureau** policy, this includes **content review**.
- For state data approved under Option A (see **??**), when reporting results from studies that include multiple states, the results should be pooled across the states. State-specific controls can be included, but no coefficients therefrom reported. Review is only subject to the usual Census Bureau rules.
- To insure that state level patterns cannot be inferred from the results, the general guidance is that projects should plan on using data from at least three states. It is also strongly encouraged that no one state represent half or more of the employment in the pooled sample (e.g. pooling Texas, Vermont, and Wyoming is not acceptable).

The identity of the LED member states is obviously not confidential (see `https://lehd.ces.census.gov/state_partners/`). You may say which states were used in your analysis, and that you controlled for state-specific factors. The chief disclosure officer for the RDC network will review compliance with this requirement in consultation with the Assistant Division Chief for LEHD.

### 1.2.1 Additional rules

Additional rules may apply to the use of the ICF (chapter 5). Please see Section 5.1.3 for more information.

## 1.3 TREATMENT OF FEDERAL TAX INFORMATION (FTI)

Some components of the LEHD Infrastructure include Title-26 protected variables. In the Snapshot, these are stored as separate datasets for tracking and monitoring purposes,and are subject to distinct permissions. This document provides information on all LEHD data, yet T26 components need to be requested separately, and as of the writing of this documentation, will trigger additional proposal review. Table 1.2 shows these components and their Federal Tax Information (FTI) counterparts, if present, as they are available in the RDC.

## 1.4 IDENTIFIERS

In general, linkages between the different files are created using deterministic match-merge techniques. Person, firm, and establishment identifiers allow users to link all LEHD Infrastructure files.

Table 1.2: LEHD components and FTI

| Name and<br>abbreviation | Described<br>in Section | Name of<br>FTI version | CES abbreviation<br>of FTI version |
|---|---|---|---|
| Employer Characteristics File (ECF) | 2.3.8 | ECFT26 | ect |
| Employment History Files (EHF) | n.a. | n.a. | n.a. |
| Individual Characteristics File (ICF) | 5.3.10 | ICFT26 | ict |
| Geocoded Address List (GAL) | 4.4.4 | GALT26 | gat |
| Office of Personnel Management (OPM) | 6.3.3.1 | OPMT26 | n.a. |
| Quarterly Workforce Indicators (QWI)<br>(establishment level) | n.a. | n.a. | n.a. |
| Successor-Predecessor File (SPF) | n.a. | n.a. | n.a. |
| Unit-to-Worker Impute (U2W) | n.a. | n.a. | n.a. |

### 1.4.1 Person identifiers

Throughout, all Social Security Numbers (SSNs) have been replaced by Protected Identity Keys (PIKs) - no SSNs are available anywhere in these data. Linkage to other person-level data products at the Census Bureau require crosswalks keyed to the PIK, which are not available as part of the LEHD Snapshot and must be requested separately.

### 1.4.2 Firm identifiers

Several firm identifiers are available. **State employer identification numbers (SEINs)** are constructed internally by LEHD, and generally, but not always, reflect an entity reporting UI taxes to state authorities. "Establishments" (more precisely: reporting units) are identified by **SEIN reporting unit (SEINUNIT)**. Establishments and firms are structured as one would expect with establishments listed hierarchically within each firm. Therefore to uniquely identify an establishment both the SEIN and SEINUNIT must be used. The firm and establishment identifiers are state and firm-structure-specific - within the LEHD Infrastructure files, there is no straighforward method of linking units of a firm with multiple tax reporting entities (SEINs). Although the vast majority of firms have only one SEIN, a firm, depending on its structure may have multiple SEINs operating both within and across state boundaries.

Although the **federal Employer Identification Number (EIN)** is available and can be used to link SEINs within and across states, the EIN suffers from similar problems as the SEIN. The identifier is not necessarily unique within a firm, is designed for tax reporting, and the structure of EINs within a firm is arbitrary. The Census Bureau recognizes the limitations of administrative identifiers and has addressed this problem on the Business Register (BR) and the Longitudinal Business Database (LBD). The Business Register Bridge (BRB) files as well as the EIN stored on the ECF are used to link to the Business Register (BR), Longitudinal Business Database (LBD) and other Census economic data. Note that the BRB is in general a many-to-many link file. The BRB does permit assigning all SEINs and SEINUNITs to a common **alpha** (the overall firm identifier in the BR). However, exact identifier-based establishment-to-establishment matches between BR/LBD and LEHD data are generally not possible for establishments part of multi-establishment firms.

**Firm or establishment names** are not available to researchers on LEHD files. Researchers who need to link firm or establishment data by name must use other datasets to perform the linkage, such as BR or Standard Statistical Establishment List (SSEL), then subsequently link to the LEHD data as outlined above.

For any further information, refer to the component-specific documentation.

## 1.5 PROCESSING FILES

LEHD Infrastructure files are significantly larger than even traditionally large research files such as the decennial census. Careful planning is required to ensure that adequate resources are available. To facilitate researchers in this endeavor, the research versions of the LEHD Infrastructure files in the Federal Statistical Research Data Center (FSRDC) environment have additional random variables that allow for the selection of uniform random subsamples of firms (SEIN), establishments (SEINUNIT), and individuals (PIK). No such random variable is available on the EHF, since there is no single good strategy for selecting jobs. Tables in the documentation for individual components also contain information about the size on-disk of each file.

## 1.6 FILE LOCATIONS

The Census Bureau occasionally updates its storage infrastructure. In this document, all file locations are relative to a generic "Snapshot data root" directory, which remains unspecified. Two directory naming conventions are in effect during the release of this snapshot, for file locations relative to the Snapshot data root (here for the example of the ECF, note the underscore):

```
ecf/ZZ/       old convention
ecf_ZZ/       new convention
```

Thus, the file `ecf_al_seinunit.sas7bdat` (Alabama ECF) may be found either at

```
(path to Snapshot data root)/ecf/al/ecf_al_seinunit.sas7bdat
```

or at

```
(path to Snapshot data root)/ecf_al/ecf_al_seinunit.sas7bdat
```

In this document, we will consistently use the old convention (**`ecf/ZZ`**).

## 1.7 CITING THE DATA AND SPONSORS

### Sponsors

The LEHD Snapshot draws on a data infrastructure that received substantial funding from a number of funding agencies and foundations. We strongly encourage researchers to acknowledge that funding in their paper's "Acknowledgements" or data appendix. The following statement can be used:

```
This research uses data from the Census Bureau's Longitudinal Employer
Household Dynamics Program, which was partially supported by the following
National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889;
National Institute on Aging Grant AG018854; and grants from the Alfred
P. Sloan Foundation.
```

### Data access

In addition, as more and more journals and funding agencies have stringent data availability requirements (National Science Foundation 2011; American Economic Association 2014; Review of Economics and Statistics 2014; Journal of Labor Economics 2009), researchers will need to work with the Census Bureau to ensure availability of their programs and research extracts. The following statement has been successfully used for accepted papers (provided by John M. Abowd, Cornell University):

> The data used for this paper were prepared in the U.S. Census Bureau’s secure computing facilities under an authorized project using the Research Data Center network. The exact analysis files have been fully archived so that the programming sequence submitted in compliance with the [JOURNAL]’s editorial policy can be run in its entirety, except for the component that extracts the analysis sample from the underlying confidential databases. I grant any researchers with appropriate Census-approved project permission to use my exact research files provided that those files were among the ones that they requested when the approval was obtained (a Census Bureau requirement). In compliance with the [JOURNAL]’s editorial policy, I am submitting the list of those files, and the last known location of the archive on the Census Bureau’s RDC network as of [date]. I authorize the editorial staff of the [JOURNAL] to release this list and my statement of cooperation to any researcher who requests it, as well as to the U.S. Census Bureau or any agency cooperating with the Census Bureau in supervising research that uses the restricted-access data that I have used.

## Data citation

A suggested data citation for each component of the LEHD Snapshot is provided in each chapter, and can be used in the bibliography of researchers’ articles (see https://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/citations.jsp for more details on data citations), for instance:

U.S. Census Bureau. 2016. *Individual Characteristics Files (ICF) in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

The full Bibtex file underlying the data citations is attached to this document. LaTeX users can simply add the bibliography file to their sources, and cite the data in the text, as they would regular articles:

```
...
I am using the S2014 ICF \citep{S2014:icf}.
...
\bibliographystyle{chicago}
\bibliography{myfile.bib,data.bib}
...
```

which would yield

```
...
I am using the S2014 ICF (U.S.  Census Bureau
2016).
...
```

**Bibliography**

U.S. Census Bureau. 2016. *Individual Characteristics Files (ICF) in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

Users of other bibliographical software can generally import Bibtex files, and should refer to their user manual.

## 1.8 CHANGES TO SNAPSHOT OVER TIME AND RELEASE HISTORY

### 1.8.1 Scope

The S2014 snapshot covers all states for which data was available at the time that the snapshot was created, as well as OPM data on covered federal workers. This snapshot extends the available time series through **2015Q1**, where possible.

### 1.8.2 Changes on the ICF

**Updated data** There has been no change to the imputation models that complete the data in the ICF. New workers having entered the LEHD covered workforce in the time period covered have been added, and where necessary, lookups and imputations for complete demographic details performed as outlined in the main document.

### 1.8.3 Changes to EHF

**New National Indicator of Employment** A data file indicating whether a worker had covered employment in any state has been added (EHF US Indicators). This file is useful for researchers who only have access to a subset of the national data, but need to identify labor force status in other states, for instance to do a Heckman-style correction. This file is designed to permit such analyses. The file is not subject to state-specific restrictions, and is automatically added to any project requesting LEHD data.

**Construction** For each PIK $i$, in each EHF state file $s$, identify the quarters $q$ in which at least $1 was earned in any job $j$:

$$\forall i, \forall q, \forall s : pos\_earn_{isq} = 1 \text{ if } \left( \sum_{\substack{j:J(i,q)=1 \\ \wedge J(s,q)=1}} earn_{ijq} \right) \geq 1$$

where $J(i,q)$ indicates that $i$ worked for $j$ in $q$ and $J(s,q)$ indicates that $j$ is located in $s$ in $q$. Then for each quarter, count the number of states for which the worker has positive earnings:

$$\forall i, \forall q : pos\_earn_{iq} = \sum_{s} pos\_earn_{isq}$$

The resulting file has a PIK-year structure with `pos_earn1` -`pos_earn4` variables indicating the number of states in which the worker had jobs in a particular year and quarter. Separately, a SAS data file with the temporal availability of states is available in the same directory (`ehf_all_availability`), so that users can control for the at-risk states at each point in time (note that this information corresponds to the metadata already made available within this document, but is more convenient in that particular format).

**EHF availability metadata** Researchers often request, or otherwise re-compute information for when each state EHF file starts and ends. To provide researchers with this information in an easy-to-use format, the file `ehf_by_state` is made available, containing start and end dates of EHF data, by state.

**File format of EHF-PHF** The EHF-PHF, a wide version of the regular EHF (one observation for job, rather than one observation per job-year), is often used by researchers since it provides convenient access to all earnings on a given job. No additional information is available on the EHF-PHF, only the organization of the data is different. In this snapshot, in order to deduplicate data and conserve scarce disk space, this file has been converted from a physical file to a SAS view. Users should notice no difference in file access speed or use of the file. Users should read the SAS documentation to learn more about SAS views.

### 1.8.4 Changes to the ECF

**Restructured ECF** The ECF has been restructured. The fundamental variables typically used by researchers remain available, but some variables have changed names or are no longer produced by the production system. Tables 2.1 on page 2-1 and 2.2 on page 2-3 provide information on these changes. **No changes have been made to T26 variables.**[2]

One of the key changes made by LEHD Production has been the updating of the possible industry codings. NAICS2012 codings are now available. SIC1987-based codings are still available, and are consistently named with the NAICS-based codings. For more information on how industry coding is propagated (deterministically or probabilistically imputed), users should consult Vilhuber and McKinney (2014). The `SRC` variables for each of the industry codings are no longer available. To identify the industry as coded on the original ES202 record, if so desired, the user should consult `QCEW_NAICS` (with associated inferred revisions of NAICS stored in `QCEW_NAICS_YR`), and `QCEW_SIC`.

### 1.8.5 Changes to GAL

The GAL is under continuous development and improvement. In this snapshot, the GAL has been completely restructured. Table 4.1 provides an overview of the available files. Due to fundamental changes to the structure of the GAL, an attempt was made to maintain backward compatibility, but not all changes could be reasonably mapped. Crosswalks to all input sources have been re-introduced. Access to certain crosswalks may require additional permissions or may not be available to RDC-based researchers due to legal or contractual restrictions.

### 1.8.6 Changes to OPM data on Federal workers

OPM data on Federal workers were first added to the Snapshot with S2011. In the S2014 version of the snapshot, OPM data have been collected under a single directory. RDC users should be able to access these files by requesting a “OPM” dataset. Access to the OPM data do not require state permissions.

### 1.8.7 Changes to QWI establishment files

The QWI_SEINUNIT files (internally known as UFF_B) have been modified since the previous snapshot, reflecting changes in the QWI publications for which they serve as inputs. Each file contains the statistics known from the public-use QWI. Variable names have changed. The QWI_SEINUNIT files contain no FTI.

#### 1.8.7.1 Restrictions

Note that the use of the QWI_SEINUNIT files is incompatible with the use of the QWI public-use files also available in the FSRDC. Researchers must choose one or the other.

#### 1.8.7.2 Changes to names of files

Prior to S2011, only *age x sex* tabulations were available, and the files were simply called “QWI_SEINUNIT”. In S2011, race, ethnicity, and education tabulations were added. In this release, two file names have been modified:

- QWI_SEINUNIT_SA is the new name of QWI_SEINUNIT_WIA, containing *age x sex* statistics
- QWI_SEINUNIT_D is the new name of QWI_SEINUNIT_estabtots, containing only the marginal categories (i.e., no breakouts by demographic specific groups)

In addition, the marginal categories contained in QWI_SEINUNIT_D are no longer duplicated in the other files.

[2] For users familiar with the LEHD production files, please note that some T26 variables that are available on LEHD production files have been dropped, as they provide redundant information to already existing T26 variables, and are present primarily for production QA.

#### 1.8.7.3 Renamed and dropped variables

In line with changed publication of QWI (lehd.ces.census.gov/doc/Memo_changes_to_QWI.pdf), several variables are no longer present, or have been renamed. In general, names are consistent with the "Alternate names" in LEHD Schema V4.0.1 (see lehd.ces.census.gov/data/schema/V4.0.1).

The following variables are no longer available:

- All variables related to *changes in total earnings* (starting with `dW`)
- All variables that are rates (ending in `R`)
- All variables related to periods of non-employment preceding or following a transition (starting with `N`)
- Certain average earnings variables (`WCA, WCS, WA, WS`)
- Certain variables relating to continuous quarter hiring (`CH`, `CR`)
- The variable `FSnx` (Full-quarter separations in the next quarter), due to a processing change (CHECK)
- The variable `W2` (average earnings for end-of-quarter employment), replaced by `W2B` (average earnings for beginning-of-quarter employment)

The following variables have been renamed:

- The name of the SIC variable has changed from `ES_SIC` to `SIC1987FNL`
- The name of the NAICS variable has changed from `ES_NAICS_FNL2007` to `NAICS2012fnl` and reflects NAICS 2012 coding.
- The name of `H3` (New Hires into Full-Quarter Employment) has been changed to `FH` for consistency with the public-use variables.
- Earnings-related variables have been made consistent with the public-use variables

| Public-use name | Internal name | Label |
|---|---|---|
| ZW3 | W3 | Average Monthly Earnings (Full-Quarter Employment) |
| ZW1 | W2B | Average Monthly Earnings (Beginning-of-Quarter Employment) |
| ZWFA | WFA | Average Monthly Earnings (All Hires into Full-Quarter Employment) |
| ZWFH | WH3 | Average Monthly Earnings (New Hires into Full-Quarter Employment) |
| ZWFS | WFS | Average Monthly Earnings (Flows out of Full-Quarter Employment) |

#### 1.8.7.4 Changes in coding (variable names)

Please note that coding for race and ethnicity has changed, see Table 7.1. This affects the naming of variables.

### 1.8.8 Changes to Successor-Predecessor File

None.

### 1.8.9 Updates: April 2013: S2011 release

The S2011 was the third release of the LEHD Infrastructure files. It contains data for the period through the end of 2011, and includes Q1 of 2012. The data was pulled from LEHD archives as a coherent ensemble in 2012Q4 and 2013Q1. The LEHD Snapshot S2011 covers 49 states and the District of Columbia. Massachusetts, the Virgin Islands, and Puerto Rico have not yet had infrastructure files produced.

We should highlight the fact that not all states have full-quality data through Q1 of 2010. Problematic interior quarters or lower-quality variables will generally be included in the Snapshot and are highlighted in their respective sections (in particular EHF and ECF) and through appropriate data quality flags. States with recent data delivery or quality issues may have shorter time series overall (data may end earlier than 2010Q1).

Public-Use QWI (QWIPU) are available for the first time in the S2011 snapshot. Note that use of the QWIPU data precludes access to the confidential files, but has certain other advantages.

For detailed information, see Vilhuber and McKinney (2014).

### 1.8.10 October 2010: S2008 release

The S2008 release is the second release of the LEHD Infrastructure files. It contains data that covers the years up to and including 2008Q1. The data was pulled from LEHD archives as a coherent ensemble in October 2009. For detailed information, see McKinney and Vilhuber (2011a).

### 1.8.11 August 2008: S2004 release

The S2004 snapshot is the first release of the LEHD Infrastructure files. It contains data that covers the years up to and including 2004Q1. The data was pulled from LEHD archives as a coherent ensemble over the course of 2005 and 2006. For detailed information, see McKinney and Vilhuber (2011b).

# Chapter 2. Employer Characteristics File (ECF)

## 2.1 OVERVIEW

The Employer Characteristics File (ECF) consolidates LEHD employer microdata information on size, location, industry, etc., into two easily accessible files. For each firm identified by SEIN, establishment-level data, identified by SEIN-SEINUNIT, is stored in the "SEINUNIT file." Some information is aggregated to the SEIN level, and stored in the "SEIN file." The SEIN file contains no new information, and should be viewed merely as an easier and/or more efficient way of accessing data aggregated to the firm level. Each file contains one record for every YEAR QUARTER a firm and/or establishment is present in either the ES-202 or the UI. All information is subject to extensive data edits and imputation, and the final files contain no missing information. For a more extensive description of data processing in the ECF, we refer the reader to Vilhuber and McKinney (2014). The files can be linked to other Census data through the use of the LEHD SEIN as well as the EIN.

### 2.1.1 Changes in Snapshot S2014

**Restructured ECF** The ECF has been restructured. The fundamental variables typically used by researchers remain available, but some variables have changed names or are no longer produced by the production system. Tables 2.1 and 2.2 on page 2-3 provide information on these changes. **No changes have been made to T26 variables**.[1]

One of the key changes made by LEHD Production has been the updating of the possible industry codings. NAICS2012 codings are now available. SIC1987-based codings are still available, and are consistently named with the NAICS-based codings. For more information on how industry coding is propagated (deterministically or probabilistically imputed), users should consult Vilhuber and McKinney (2014). The `SRC` variables for each of the industry codings are no longer available. To identify the industry as coded on the original ES202 record, if so desired, the user should consult `QCEW_NAICS` (with associated inferred revisions of NAICS stored in `QCEW_NAICS_YR`), and `QCEW_SIC`.

Table 2.1: Dropped and added variables on S2014 ECF

| No longer available in S2014 | New in S2014 |
|---|---|
| *SEINUNIT file* | |
| es_county_miss | leg_block_src |
| es_naics_fnl1997_miss | |
| es_naics_fnl2002_miss | qcew_naics |
| es_naics_fnl2007_miss | qcew_naics_aux |

[1] For users familiar with the LEHD production files, please note that some T26 variables that are available on LEHD production files have been dropped, as they provide redundant information to already existing T26 variables, and are present primarily for production QA.

Table 2.1 (cont.): Dropped and added variables

| No longer available in S2014 | New in S2014 |
|---|---|
| es_owner_code_miss | qcew_naics_aux_yr |
| leg_county_orig | qcew_naics_yr |
| leg_galid_orig | es_naics_fnl2012 |
| leg_geo_qual_orig | |
| multi_unit | |
| yr_qtr | |
| ES_NAICS_FNL1997_SRC | |
| ES_NAICS_FNL2002_SRC | |
| ES_NAICS_FNL2007_SRC | |
| ES_SIC_DIV | |
| ES_SIC_SRC | |
| es_sic_miss | |
| *SEIN file* | |
| mode_es_naics_fnl1997 | mode_naics2012fnl_emp |
| mode_es_naics_fnl2002 | ui_payroll |
| mode_es_naics_fnl2007 | qtime |
| mode_es_county | multi_first_qtime |
| mode_es_owner_code | ehf_source |
| mode_es_sic | ehf_sourcetp |
| mode_leg_cbsa | es_state |
| mode_leg_cbsa_memi | |
| mode_leg_county | |
| mode_leg_county_orig | |
| mode_leg_county_orig_emp | |
| mode_leg_state | |
| mode_leg_state_emp | |
| mode_leg_subctygeo | |
| mode_leg_wib | |
| mode_es_county_emp_miss | |
| mode_es_county_miss | |
| mode_es_naics_fnl1997_emp_miss | |
| mode_es_naics_fnl1997_miss | |
| mode_es_naics_fnl2002_emp_miss | |

Table 2.1 (cont.): Dropped and added variables

| No longer available in S2014 | New in S2014 |
|---|---|
| mode_es_naics_fnl2002_miss | |
| mode_es_naics_fnl2007_emp_miss | |
| mode_es_naics_fnl2007_miss | |
| mode_es_owner_code_emp_miss | |
| mode_es_owner_code_miss | |
| mode_es_sic_emp_miss | |
| mode_es_sic_miss | |
| mode_leg_cbsa_emp_miss | |
| mode_leg_cbsa_memi_emp | |
| mode_leg_cbsa_miss | |
| mode_leg_county_emp_miss | |
| mode_leg_county_miss | |
| mode_leg_county_orig_emp_miss | |
| mode_leg_county_orig_miss | |
| mode_leg_state_emp_miss | |
| mode_leg_state_miss | |
| mode_leg_subctygeo_emp_miss | |
| mode_leg_subctygeo_miss | |
| mode_leg_wib_emp_miss | |
| mode_leg_wib_miss | |
| multi_unit_code | |

Table 2.2: Name changes, ECF

| **S2011** | **S2014** |
|---|---|
| *SEINUNIT file* | |
| source | ehf_source |
| *SEINUNIT file* | |
| mode_es_naics_fnl1997_emp | mode_naics1997fnl_emp |
| mode_es_naics_fnl2002_emp | mode_naics2002fnl_emp |
| mode_es_naics_fnl2007_emp | mode_naics2007fnl_emp |
| mode_es_sic_emp | mode_sic1987fnl_emp |

## 2.2 DATA CITATION

U.S. Census Bureau. 2016. *Employer Characteristics Files (ECF) in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

## 2.3 DATA SET DESCRIPTIONS

### 2.3.1 Naming scheme

There are five files in the ECF group, and one additional file in the ECFT26 group:

- ecf/zz/ecf_zz_sein_aux.sas7bdat
- ecf/zz/ecf_zz_sein.sas7bdat
- ecf/zz/ecf_zz_seinunit_aux.sas7bdat
- ecf/zz/ecf_zz_seinunit_bii.sas7bdat
- ecf/zz/ecf_zz_seinunit.sas7bdat
- ecft26/zz/ecf_zz_t26.sas7bdat

ZZ stands for the state postal abbreviation. Files with _bii contain state-provided Business-identifying information (BII), and while they do not require special permissions, they may be deleted under certain circumstances. Files with _t26 contain FTI, are stored in separate subdirectories and require a separate set of permissions. Files with _cc or _fuzz contain Census-confidential information and are generally not available to external researchers. Either set of files are of little use without the regular ECF group data.

The ECF files can be large, and researchers may wish to analyze only a random subsample of firms. The variables SAMPLE_SEIN and SAMPLE_SEINUNIT can be used to select a random sample of the ECF. For details, see the programs in Section 2.4.

### 2.3.2 Data location

The files are stored in three main directories, with state-specific subdirectories:

```
ecf/ZZ/       for most files
ecft26/ZZ     for files with Title 26 protected content
```

### 2.3.3 Main SEINUNIT dataset: ecf_zz_seinunit

SEINUNIT-level file, research variables only.

**Record identifier:** SEIN SEINUNIT YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER SEINUNIT
**File indexes:** none
**Entity** "establishment" or State Employment Security Agency (SESA)
**Unique Entity Key** SEIN SEINUNIT

Note that SEINUNIT is only unique within any given time period within SEIN.

Table 2.3: ECF_ZZ_SEINUNIT: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **best_emp1** | Best UI/202 Employment Month 1 | Num | 8 |
| **best_emp2** | Best UI/202 Employment Month 2 | Num | 8 |
| **best_emp3** | Best UI/202 Employment Month 3 | Num | 8 |
| **best_flag** | Source of best_ data | Num | 8 |
| **best_wages** | Best UI/202 Wages | Num | 8 |
| **es_county** | ES202 County code, edited | Char | 3 |
| **es_naics_fnl1997** | Final NAICS 1997 code | Char | 6 |
| **es_naics_fnl2002** | Final NAICS 2002 code | Char | 6 |
| **es_naics_fnl2007** | Final NAICS 2007 code | Char | 6 |
| **es_naics_fnl2012** | Final NAICS 2012 code | Char | 6 |
| **es_owner_code** | Owner code, edited | Char | 1 |
| **es_sic** | Final SIC 1987 code | Char | 6 |
| **es_state** | ES202 FIPS State SS | Char | 2 |
| **leg_block** | Census Block | Char | 4 |
| **leg_block_src** | Code1; Maf; Equals maf; btWn maf; cOmmercial; mIxed; reSidential; all aDdresses | Char | 1 |
| **leg_block_suf1** | Census Block suffix 1 | Char | 1 |
| **leg_block_suf2** | Census Block suffix 2 | Char | 1 |
| **leg_cbsa** | Core-Based Statistical Area | Char | 5 |
| **leg_cbsa_memi** | CBSA type 1=Metro, 2=Micro, else=9 | Char | 1 |
| **leg_county** | Edited county code, final | Char | 5 |
| **leg_galid** | Final GALID | Char | 29 |
| **leg_geo_qual** | Quality of final geography | Num | 3 |
| **leg_geocode** | FIPS Tab State\|\|FIPS Tab County\|\|Census Tract | Char | 11 |
| **leg_latitude** | Latitude, 6 implied decimal places | Num | 8 |
| **leg_longitude** | Longitude, 6 implied decimal places | Num | 8 |

(cont.)

**Table 2.3 (cont.): ECF_ZZ_SEINUNIT: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **leg_state** | Cleaned GEO State SS | Char | 5 |
| **leg_subctygeo** | Sub-county geography from LEG | Char | 10 |
| **leg_wib** | Workforce Investment Board area | Char | 6 |
| **num_estabs** | Number of Establishments | Num | 4 |
| **qcew_auxiliary_code** | Auxiliary Industry Code | Char | 1 |
| **qtime** | Quarter index (1985Q1=1) | Num | 3 |
| **quarter** | Quarter (Numeric) | Num | 3 |
| **sample_sein** | Random sample selector for SEIN | Num | 8 |
| **sample_seinunit** | Random sample selector for SEINUNIT | Num | 8 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State establishment identifier | Char | 5 |
| **year** | Year (YYYY) | Num | 3 |

### 2.3.4 Auxiliary SEINUNIT dataset: ecf_zz_seinunit_aux

SEINUNIT-level file, auxiliary and diagnostic variables only.

**Record identifier:** SEIN SEINUNIT YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER SEINUNIT
**File indexes:** none
**Entity** "establishment" or SESA
**Unique Entity Key** SEIN SEINUNIT

This file can be merged onto the main SEINUNIT file using the specified identifiers in sort order. It is generally not needed by researchers, but made available for those requiring more detailed longitudinal information on imputes and edits.

Table 2.4: ECF_ZZ_SEINUNIT_AUX: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **base_ind_code_year** | Latest coding year in reported industry data | Num | 3 |
| **ehf_source** | EHF: Source of Earnings data (FIPS state code/0=Fed) | Char | 2 |
| **ehf_sourcetp** | EHF: Type of source for earnings data | Char | 2 |
| **es_county_flag** | County longitudinal edit flag | Num | 3 |
| **es_galid** | GALID of address on es202 | Char | 29 |
| **es_geo_qual** | Quality code for geography from es202 | Num | 3 |
| **es_naics1997init** | Initial NAICS 1997 code | Char | 6 |
| **es_naics2002init** | Initial NAICS 2002 code | Char | 6 |
| **es_naics2007init** | Initial NAICS 2007 code | Char | 6 |
| **es_naics2012init** | Initial NAICS 2012 code | Char | 6 |
| **es_naics_yr** | NAICS coding year (implied) | Num | 3 |
| **es_owner_code_flag** | Owner code longitudinal edit flag | Num | 3 |
| **es_owner_code_flag_init** | Owner Code Flag | Num | 3 |
| **es_owner_code_init** | ES202 Owner code, invalid values removed | Char | 1 |
| **industry_source** | Industry source flag | Char | 1 |
| **ldb_naics** | NAICS code reported on LDB | Char | 6 |
| **ldb_naics_flag** | Flag for missing NAICS_LDB | Num | 3 |
| **ldb_naics_yr** | NAICS_LDB coding year (implied) | Num | 3 |
| **move** | LEG, Flag indicating seinunit move | Num | 3 |
| **qcew_county** | County edited to remove those invalid in current geography | Char | 3 |
| **qcew_county_flag** | Flag for missing county | Num | 3 |
| **qcew_county_raw** | County as reported on ES202 | Char | 3 |
| **qcew_empl_month1** | ES202 Employment Month 1, edited | Num | 5 |
| **qcew_empl_month2** | ES202 Employment Month 2, edited | Num | 5 |

(cont.)

**Table 2.4 (cont.): ECF_ZZ_SEINUNIT_AUX: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **qcew_empl_month3** | ES202 Employment Month 3, edited | Num | 5 |
| **qcew_empl_month1_flg** | Reported or imputed Month 1 Employment | Char | 1 |
| **qcew_empl_month2_flg** | Reported or imputed Month 2 Employment | Char | 1 |
| **qcew_empl_month3_flg** | Reported or imputed Month 3 Employment | Char | 1 |
| **qcew_naics** | NAICS code reported on ES202 | Char | 6 |
| **qcew_naics_aux** | Auxiliary NAICS code reported on ES202 | Char | 6 |
| **qcew_naics_aux_flag** | Flag for missing NAICS_AUX | Num | 3 |
| **qcew_naics_aux_yr** | NAICS_AUX coding year (implied) | Num | 3 |
| **qcew_naics_flag** | Flag for missing NAICS | Num | 3 |
| **qcew_sic** | SIC as reported on ES202 | Char | 4 |
| **qcew_sic_flag** | Flag for missing SIC | Num | 3 |
| **qcew_sic_yr** | SIC coding year (implied) | Num | 3 |
| **qcew_total_wages** | ES202 wages, edited | Num | 7 |
| **qcew_total_wages_flg** | Reported or imputed Total Wages | Char | 1 |
| **quarter** | Quarter (Numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State establishment identifier | Char | 5 |
| **year** | Year (YYYY) | Num | 3 |

### 2.3.5 Main SEIN dataset: ecf_zz_sein

SEIN-level file, with variables aggregated from the establishment level.

**Record identifier:** SEIN YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER
**File indexes:** none
**Entity** “firm”
**Unique Entity Key** SEIN

Note that SEIN is unique within any given time period across all states, but may not be uniquely identify an entity over time within a state, as the underlying UI account numbers can and do get re-used.

Table 2.5: ECF_ZZ_SEIN: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **ehf_source** | EHF: Source of Earnings data (FIPS state code/0=Fed) | Char | 2 |
| **ehf_sourcetp** | EHF: Type of source for earnings data | Char | 2 |
| **es_state** | ES202 FIPS State SS | Char | 2 |
| **mode_es_county_emp** | Employment mode ES202 County code | Char | 3 |
| **mode_es_naics_fnl1997_emp** | Employment mode Final NAICS 1997 code | Char | 6 |
| **mode_es_naics_fnl2002_emp** | Employment mode Final NAICS 2002 code | Char | 6 |
| **mode_es_naics_fnl2007_emp** | Employment mode Final NAICS 2007 code | Char | 6 |
| **mode_es_naics_fnl2012_emp** | Employment mode Final NAICS 2012 code | Char | 6 |
| **mode_es_owner_code_emp** | Employment mode owner code | Char | 1 |
| **mode_es_sic_emp** | Employment mode Final SIC 1987 code | Char | 6 |
| **mode_leg_cbsa_emp** | Employment mode Core-based statistical area | Char | 5 |
| **mode_leg_county_emp** | Employment mode edited county code | Char | 5 |
| **mode_leg_subctygeo_emp** | Employment mode sub-county geography code | Char | 10 |
| **mode_leg_wib_emp** | Employment mode Workforce Investment Board area | Char | 6 |
| **multi_first_qtime** | First qtime SEIN reports as multi-establishment | Num | 3 |
| **multi_first_quarter** | First quarter SEIN reports as multi-establishment | Num | 3 |
| **multi_first_year** | First year SEIN reports as multi-establishment | Num | 3 |
| **multi_unit** | SEIN with 2 or more establishments on 202 | Num | 3 |
| **num_estabs** | Number of Establishments | Num | 4 |
| **qtime** | Quarter index (1985Q1=1) | Num | 3 |
| **quarter** | Quarter (Numeric) | Num | 3 |
| **qwi_unit_weight** | Weight sum(B_UI)=sum(month1_BLS) | Num | 8 |
| **sample_sein** | Random sample selector for SEIN | Num | 8 |
| **sein** | State Employer Identification Number | Char | 12 |

(cont.)

**Table 2.5 (cont.): ECF_ZZ_SEIN: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **sein_best_emp1** | SEIN Best UI/202 Month 1, Employment | Num | 8 |
| **sein_best_emp2** | SEIN Best UI/202 Month 2, Employment | Num | 8 |
| **sein_best_emp3** | SEIN Best UI/202 Month 3, Employment | Num | 8 |
| **sein_best_wages** | SEIN Best UI/202 Payroll | Num | 8 |
| **ui_payroll** | Original UI Payroll Info W1 | Num | 8 |
| **year** | Year (YYYY) | Num | 3 |

### 2.3.6 Auxiliary SEIN dataset: ecf_zz_sein_aux

SEIN-level file, auxiliary and diagnostic variables only.

**Record identifier:** SEIN YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER
**File indexes:** none
**Entity** “firm”
**Unique Entity Key** SEIN

This file can be merged onto the main SEIN file using the specified identifiers in sort order. It is generally not needed by researchers, but made available for those requiring more detailed longitudinal information on imputes and edits.

Table 2.6: ECF_ZZ_SEIN_AUX: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **ever_202** | SEIN ever on 202 | Num | 3 |
| **ever_UI** | SEIN ever on UI | Num | 3 |
| **ever_multi** | SEIN ever multi unit | Num | 3 |
| **in_202** | SEIN in ES202 | Num | 3 |
| **in_UI** | SEIN in UI | Num | 3 |
| **mode_es_county_flag** | Employment mode ES202 County code longitudinal edit flag | Num | 8 |
| **mode_es_owner_code_flag** | Employment mode owner code longitudinal edit flag | Num | 8 |
| **qcew_sein_emp1** | SEIN 202 Employment Month 1 | Num | 8 |
| **qcew_sein_emp2** | SEIN 202 Employment Month 2 | Num | 8 |
| **qcew_sein_emp3** | SEIN 202 Employment Month 3 | Num | 8 |
| **qcew_sein_wages** | SEIN 202 Wages | Num | 8 |
| **quarter** | Quarter (Numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **ui_seinsize_b** | UI Employment B | Num | 8 |
| **ui_seinsize_e** | UI Employment E | Num | 8 |
| **ui_seinsize_m** | UI Employment M | Num | 8 |
| **year** | Year (YYYY) | Num | 3 |

### 2.3.7 Auxiliary BII dataset: ecf_zz_seinunit_bii

State-provided Business-identifying information (BII) is segregated into this file. Since name information is not retained by LEHD for processing on the ECF, and address information is only collected by the GAL, this primarily pertains to information regarding the original state-provided, but federally-sourced Employer Identification Number (EIN). For California only, the EIN-related information is on the ECF T26, and absent from this file.

**Record identifier:** SEIN SEINUNIT YEAR QUARTER
**Sort order:** SEIN SEINUNIT YEAR QUARTER
**File indexes:** none
**Entity** "establishment" or SESA
**Unique Entity Key** SEIN SEINUNIT

Table 2.7: ECF_ZZ_SEINUNIT_BII: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **ein** | ES-202 EIN (Clean) | Char | 9 |
| **qcew_ein** | Raw (unmodified ES202 EIN) | Char | 9 |
| **qcew_ein_defect** | Flag for defect type on ein | Num | 3 |
| **quarter** | Quarter (Numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State establishment identifier | Char | 5 |
| **ui_ein** | modal EIN from UIPIK data | Char | 9 |
| **ui_ein_defect** | Flag for defect type on ui_ein | Num | 8 |
| **ui_ein_flag** | Quarters Away UI_EIN data found | Num | 4 |
| **ui_ein_miss** | 0=ok,1=not found,2=found off qtr | Num | 4 |
| **valid_ein** | Flag for valid ein | Num | 3 |
| **year** | Year (YYYY) | Num | 3 |

### 2.3.8 Auxiliary T26 dataset: ecf_zz_t26

T26 variables associated with both the SEIN and the SEINUNIT-level file. For California, this includes the EIN-related variables as well. For all states, this includes any variables derived from T26 datasets, primarily the BR. National firm-age and firm-size variables are on this file.

**Record identifier:** SEIN SEINUNIT YEAR QUARTER
**Sort order:** SEIN SEINUNIT YEAR QUARTER
**File indexes:** none
**Entity** "establishment" or SESA
**Unique Entity Key** SEIN SEINUNIT

Table 2.8: ECF_ZZ_T26: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **fas_ein** | EIN cleaned for Firm Age & Size Proj (plus 00000) | Char | 14 |
| **fas_ein_flag** | Info about fas_ein variable | Num | 3 |
| **fas_ein_match_lbd** | 1=EIN matches to LBD this year,2=next year,3=previous year,4=NECF(no LBD),5=ECF only | Num | 8 |
| **fas_firm_age** | Firm Age, calculated | Num | 8 |
| **fas_firm_age_flag** | 1=Age from state ECF\|2=Age from LBD/NECF\|3=Age from earlier LDB/NECF | Num | 8 |
| **fas_firm_id** | Firm Alpha from LBD | Char | 10 |
| **fas_firm_size** | Best initial firm size (size March 12 of last year, current size if new) | Num | 8 |
| **fas_firm_size_fuzz** | Noise infused value of firmsize | Num | 8 |
| **fas_source_age** | 1=BDS,2=ECF,3=Impute | Char | 1 |
| **fas_source_size** | 1=BDS,2=ECF,3=Impute | Char | 1 |
| **multi_unit_lbd** | Multi Unit firm in LBD microdata | Num | 8 |
| **quarter** | Quarter (Numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State establishment identifier | Char | 5 |
| **year** | Year (YYYY) | Num | 3 |

## 2.4 HELPFUL PROGRAMS

The following programs might be found to be useful when using the data.

### 2.4.1 Selecting a random sample of establishments

The ECF files can be large, and researchers may wish to analyze only a random subsample of firms. The variables SAMPLE_SEIN and SAMPLE_SEINUNIT can be used to select a random sample of the ECF. To do this in a space-efficient way, the following code can be used as a template.

```
%let state=tx;

data mydata/view=mydata;
   set INLIB.ecf_&state._seinunit
       (where=(sample_seinunit <= 0.05));
run;

proc reg data=mydata;
model y= x w z;
run;
```

The code above uses a VIEW, which means the dataset is constructed on the fly as it is used in the analysis procedure. Although overall disk usage is not necessarily smaller when using random access (as the SAS regression procedure apparently does), it is still faster. For other processes using sequential access only, in particular simple DATA steps, a view will be space-efficient because only the relevant observations are streamed into any intermediate data files.

## 2.5 NOTES

# Chapter 3.
# Employment History Files (EHF)

## 3.1 OVERVIEW

The Employment History Files (EHF) are designed to store the complete in-state history of employment, for each individual that appears in the UI wage records employed at some firm, and for each firm and establishment that appear in the QCEW records with positive employment at some time.

The core EHF for each state contains one record for each employee-employer combination–a job–in that state in each year. Both annual and quarterly earnings variables are available in the EHF. Individuals who are employed, but never have strictly positive earnings at their employing SEIN (a theoretical possibility) in a given year do not have a record in the EHF for that year.

To facilitate analysis, the EHF data are restructured into "wide" file containing one observation per job (PIK-SEIN combination), with all quarterly earnings and activity information available on that record. The restructured file is called the Person History File (PHF). It should be noted that the actual file structure is at the PIK-SEIN-SEINUNIT-YEAR level for the EHF, and at the PIK-SEIN-SEINUNIT level for the PHF. Although only one state (Minnesota) has non-zero values for SEINUNIT, this allows the file structure to be homogeneous across states. An active job within a quarter, the primary job-level economic activity measure, is defined as having strictly positive quarterly earnings for the individual-employer pair that define the job.

Researchers often combine the EHF with the U2W, in order to obtain establishment-level information on jobs. The merged file, internally called PHF_b, is referred to as the Job History File (JHF) in the Snapshot. Note that whereas the LEHD production system constructs this variable in the QWI sequence, it is available in the Snapshot as part of the EHF files.

A history of observed activity (positive employment) in the QCEW records, is available and computed at the SEINUNIT level (Unit History File, UHF) and the SEIN level (SEIN History File, SHF).

### 3.1.1 Changes in Snapshot S2014

**New National Indicator of Employment** A data file indicating whether a worker had covered employment in any state has been added (EHF US Indicators). This file is useful for researchers who only have access to a subset of the national data, but need to identify labor force status in other states, for instance to do a Heckman-style correction. This file is designed to permit such analyses. The file is not subject to state-specific restrictions, and is automatically added to any project requesting LEHD data.

**Construction** For each PIK $i$, in each EHF state file $s$, identify the quarters $q$ in which at least $1 was earned in any job $j$:

$$\forall i, \forall q, \forall s : pos\_earn_{isq} = 1 \text{ if } \left( \sum_{\substack{j:J(i,q)=1 \\ \wedge J(s,q)=1}} earn_{ijq} \right) \geq 1$$

where $J(i,q)$ indicates that $i$ worked for $j$ in $q$ and $J(s,q)$ indicates that $j$ is located in $s$ in $q$. Then for each quarter, count the number of states for which the worker has positive earnings:

$$\forall i, \forall q : pos\_earn_{iq} = \sum_s pos\_earn_{isq}$$

The resulting file has a PIK-year structure with `pos_earn1` -`pos_earn4` variables indicating the number of states in which the worker had jobs in a particular year and quarter. Separately, a SAS data file with the temporal availability of states is available in the same directory (`ehf_all_availability`), so that users can control for the at-risk states at each point in time (note that this information corresponds to the metadata already made available within this document, but is more convenient in that particular format).

**EHF availability metadata** Researchers often request, or otherwise re-compute information for when each state EHF file starts and ends. To provide researchers with this information in an easy-to-use format, the file `ehf_by_state` is made available, containing start and end dates of EHF data, by state.

**File format of EHF-PHF** The EHF-PHF, a wide version of the regular EHF (one observation for job, rather than one observation per job-year), is often used by researchers since it provides convenient access to all earnings on a given job. No additional information is available on the EHF-PHF, only the organization of the data is different. In this snapshot, in order to deduplicate data and conserve scarce disk space, this file has been converted from a physical file to a SAS view. Users should notice no difference in file access speed or use of the file. Users should read the SAS documentation to learn more about SAS views.

## 3.2 DATA CITATION

U.S. Census Bureau. 2016. *Employment History Files (EHF) in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

## 3.3 INPUT FILES

### 3.3.1 Wage records: UI

Wage records correspond to the report of an individual's UI-covered earnings by an employing entity, identified by a state UI account number (called the SEIN in the LEHD system). An individual's UI wage record is retained in the processing if at least one employer reports earnings of at least one dollar for that individual during the quarter. Thus, an in-scope job must produce at least one dollar of UI-covered earnings during a given quarter in the LEHD universe. Maximum earnings reported are defined in a specific state's unemployment insurance system, and observed top-coding varies across states and over time.

A record is completed with information on the individual's Social Security Number (later replaced with the PIK within the LEHD system), first name, last name, and middle initial. A few states include additional information: the firm's reporting unit or establishment (SEINUNIT), available for Minnesota, and a crucial component to the Unit-to-Worker impute described later; weeks worked, available for some years in Florida; hours worked, available for Washington and Minnesota state.

Current UI wage records are reported for the quarter that ended approximately six months prior to the reporting date at Census (the first day of the calendar quarter). Wage records are also reported for the quarter that the state considers "final" in the sense that revisions to its administrative UI wage record data base after that date are relatively rare. This quarter typically ends nine months prior to the reporting date. Historical UI wage records were assembled by the partner states from their administrative record backup systems.

### 3.3.2 Employer reports: QCEW - ES-202

The employer reports are based on information from each state's Department of Employment Security. The data are collected as part of the Covered Employment and Wages (CEW) program, also known as the ES-202 (ES-202) program, which is jointly administered by the Bureau of Labor Statistics (BLS) and the Employment Security Agencies in a federal-state partnership. This cooperative program between the states and the federal government collects employment, payroll, and economic activity, and physical location information from employers covered by state unemployment insurance programs and from employers subject to the reporting requirements of the ES-202 system. The employer and work place reports from this system are the same as the data reported to the BLS as part of the QCEW, but are referred to in the LEHD system by their old acronym "ES-202." The universe for these data is a 'reporting unit,' which is the QCEW establishment–the place where the employees actually perform their work. Most employers have one establishment ('single-units'), but most employment is with employers who have multiple establishments ('multi-units'). One report per establishment per quarter is filed. These data are also used to compile the QCEW and the Business Employment Dynamics (BED) data at the BLS.

The information contained in the ES-202 reports has increased substantially over the years. Employers report wages subject to statutory payroll taxes on this form, together with some other information. Common to all years, and critical to LEHD processing, are information on the employer's identity (the SEIN), the reporting unit's identify (SEINUNIT), ownership information, employment on the 12th of each month covered by the quarter, and total wages paid over the course of the quarter. Additional information pertains to industry classifications (initially SIC, and later NAICS). Other information include the federal EIN, geography both at a high level (county or MSA) and low level (physical location street address and mailing address). A recent expansion of the standard report's record layout has increased the informational content substantially. The LEHD Infrastructure File system is, fundamentally, a job-based frame designed to be represent the universe of individual-employer pairs covered by state unemployment insurance system reporting requirements. Thus, the underlying data are wage records extracted from Unemployment Insurance (UI) administrative files from each LED partner state. In addition to the UI wage records, LED partner states also deliver an extract of the file reported to the Bureau of Labor Statistic's Quarterly Census of Employment and Wages (QCEW, formerly known as ES-202). These data are received by LEHD on a quarterly basis, with historical time series extending back to the early 1990s for many states.

## 3.4 DATA SET DESCRIPTIONS

### 3.4.1 Naming scheme

Most files stem from the LEHD production process 'ehf' and start with `ehf`. The JHF is created in the LEHD production process 'qwi' but is structurally equivalent to the EHF files, and starts with `jhf`. `ZZ` stands for the state postal abbreviation. The main EHF and the JHF files have no suffixes, other files have a suffix, explained in this section.

- ehf/zz/ehf_zz_controltotals.sas7bdat
- ehf/zz/ehf_zz_phf.sas7bdat
- ehf/zz/ehf_zz.sas7bdat
- ehf/zz/ehf_zz_sein_employment.sas7bdat
- ehf/zz/ehf_zz_shf.sas7bdat
- ehf/zz/ehf_zz_uhf.sas7bdat
- ehf/zz/ehf_zz_uniqpik.sas7bdat
- ehf/zz/jhf_zz.sas7bdat

Two additional files contain cross-state information (see Section 3.4.5):

- ehf/all/ehf_all_availability.sas7bdat
- ehf/all/ehf_us_indicators.sas7bdat

### 3.4.2 Data location

All files are stored in a main `ehf` directory, with most in state-specific subdirectories:

```
ehf/zz/ for most files
ehf/all/ for cross-state files
```

No files in the EHF process contain Title 26 data.

## 3.4.3 UI-based Output Files

### 3.4.3.1 EHF

The EHF is designed to store the complete in-state work history for each individual that appears in the UI wage records. The EHF for each state contains one record for each employee-employer combination in that state in each year. Every individual who is employed during a given year will then have one observation per employer for that year. Annual earnings and quarterly earnings variables are present on the file. The presence of positive quarterly earnings is used in the job flow analysis not only to compute earnings and payroll statistics but also to determine an individual's employment status each quarter.

The EHF (`ehf_&state.`) is organized by PIK-SEIN-SEINUNIT-YEAR. Note that all states except Minnesota (MN) have SEINUNIT='00000', so this reverts back to PIK-SEIN-YEAR for all states except MN.

**Record identifier** PIK-SEIN-SEINUNIT-YEAR<br>
**Sort order** PIK-SEIN-SEINUNIT-YEAR<br>
**Entity** Job<br>
**Unique Entity Key** PIK-SEIN-SEINUNIT

Table 3.1: EHF_ZZ: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **earn1** | Qtr 1 earnings | Num | 5 |
| **earn2** | Qtr 2 earnings | Num | 5 |
| **earn3** | Qtr 3 earnings | Num | 5 |
| **earn4** | Qtr 4 earnings | Num | 5 |
| **earn_ann** | Annual earnings | Num | 8 |
| **pik** | Protected Identification Key | Char | 9 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **source** | EHF: Source of Earnings data (FIPS state code/0=Fed) | Char | 2 |
| **sourcetp** | EHF: Type of source for earnings data | Char | 2 |
| **year** | Calendar year | Num | 3 |

#### 3.4.3.2 (proto-)PHF

The proto PHF is a reformatted version of the EHF. Rather than having one record per year, the PHF is organized by “job”, or unique employee-employer combination, identified by PIK-SEIN(-SEINUNIT), with complete historical arrays for earnings (`eNNN` variables) and employment status (`work` variable). Only *observed* SEINUNIT are used. It is not to be confused with the PHF_B of the QWI sequence (called JHF in the Snapshot, see Section 3.4.3.3), which is augmented with information from the U2W process.

The PHF (`ehf_&state._phf`) is organized by PIK-SEIN-SEINUNIT. Note that all states except MN have SEINUNIT=’00000’, so this reverts back to PIK-SEIN for all states except MN.

The PHF is stored as a SAS view.

**Record identifier** PIK-SEIN-SEINUNIT<br>
**Sort order** PIK-SEIN-SEINUNIT<br>
**Entity** Job<br>
**Unique Entity Key** PIK-SEIN-SEINUNIT

Table 3.2: EHF_PHF: Variables and Attributes

| | | | |
|---|---|---|---|
| **e21** | Employment in QTIME=21 | Num | 5 |
| **e22** | Employment in QTIME=22 | Num | 5 |
| **e23** | Employment in QTIME=23 | Num | 5 |
| **...** | | | |
| **e122** | Employment in QTIME=122 | Num | 5 |
| **e123** | Employment in QTIME=123 | Num | 5 |
| **e124** | Employment in QTIME=124 | Num | 5 |
| **flag_seinunit_imputed** | SEINUNIT imputed (never true, compatibility) | Num | 3 |
| **pik** | Protected Identification Key | Char | 9 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **work** | Binary workhistory ...00111000... 1=employed | Char | 112 |

#### 3.4.3.3 JHF

The JHF (jhf_&state.) is created by combining the U2W with the EHF_PHF. This creates a file with multiple imputed establishment assignments for each job, where establishment assignments are missing (multi-units in states other than Minnesota). Internally, this file is called PHF_B, produced by the QWI process. For observed establishments, flag_seinunit_imputed=0 and only one SEINUNIT will be observed. Otherwise, ten implicates seinunit1-seinunit10 are kept on the file.

**Record identifier** PIK-SEIN
**Sort order** PIK-SEIN
**Entity** Job
**Unique Entity Key** PIK-SEIN

Table 3.3: JHF_ZZ: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **e21** | Employment in QTIME=21 | Num | 5 |
| **...** | | | |
| **e124** | Employment in QTIME=124 | Num | 5 |
| **ehf_source** | EHF: Source of Earnings data (FIPS state code/0=Fed) | Char | 2 |
| **ehf_sourcetp** | EHF: Type of source for earnings data | Char | 2 |
| **first_acc** | First accession (First quarter of employment at this job spell) | Num | 3 |
| **flag_seinunit_imputed** | Flag: SEINUNIT is imputed=1 | Num | 3 |
| **last_sep** | Last separation (Last quarter of employment at this job spell) | Num | 3 |
| **pik** | Protected Identification Key | Char | 9 |
| **pred_qs** | Quarter of separation from predecessor | Num | 3 |
| **pred_sep_last6** | Employment indicators for last six quarters at predecessor job spell | Char | 6 |
| **random_pik_group** | Selector based on random PIK | Char | 2 |
| **sein** | State Employer Identification Number | Char | 12 |
| **sein_pred** | SEIN of predecessor | Char | 12 |
| **sein_succ** | SEIN of successor | Char | 12 |
| **seinunit1** | State UI Reporting Unit Number (Impute 1) | Char | 5 |
| **...** | | | |
| **seinunit10** | State UI Reporting Unit Number (Impute 10) | Char | 5 |
| **spell_u2w** | Spell count as per U2W | Num | 3 |
| **spell_u2w_pred** | Spell count as per U2W predecessor | Num | 8 |
| **spell_u2w_succ** | Spell count as per U2W successor | Num | 8 |
| **state** | State (FIPS) | Char | 2 |
| **succ_acc_first6** | Employment indicators for first six quarters at successor job spell | Char | 6 |
| **succ_qa** | Quarter of accession at successor | Num | 3 |

#### 3.4.3.4 UNIQPIK file

The UNIQPIK file contains a time-invariant list of PIKs as they appear on the EHF for a given state. Using the `cut` variable, which approximates a uniform distribution in 100 buckets, it can be used to select person-based samples across the Snapshot.

**Record identifier** PIK
**Sort order** PIK
**Entity** Person
**Unique Entity Key** PIK

Table 3.4: EHF_ZZ_UNIQPIK: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---:|
| **cut** | cut=substr(pik,1,2) | Char | 9 |
| **pik** | Protected Identification Key | Char | 9 |
| **ssnflag** | Illegal SSN Range Flag | Char | 1 |

#### 3.4.3.5 SEIN_EMPLOYMENT

The SEIN_EMPLOYMENT is a SEIN-level measure of employment based on UI data. No SEINUNIT version exists; for SEINUNIT-based versions of $b$, $e$, and $m$, see Section 7.3.5.

**Record identifier** SEIN-YEAR
**Sort order** SEIN-YEAR
**Entity** Firm
**Unique Entity Key** SEIN

Table 3.5: EHF_ZZ_SEIN_EMPLOYMENT: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **b** | Beginning of quarter employment | Num | 8 |
| **e** | End of quarter employment | Num | 8 |
| **m** | Flow employment | Num | 8 |
| **quarter** | Quarter | Num | 4 |
| **sein** | State Employer Identification Number | Char | 12 |
| **source** | EHF: Source of Earnings data (FIPS state code/0=Fed) | Char | 2 |
| **sourcetp** | EHF: Type of source for earnings data | Char | 2 |
| **w1** | Total earnings during the quarter | Num | 8 |
| **year** | Year | Num | 4 |
| **yr_qtr** | Year-Quarter YYYY:Q | Char | 6 |

### 3.4.4 ES202-based Output Files

#### 3.4.4.1 UHF

The UHF (Unit History File) contains a full employment history for each SEIN-SEINUNIT (wide file).

**Record identifier** SEIN-SEINUNIT
**Sort order** SEIN-SEINUNIT
**Entity** Establishment
**Unique Entity Key** SEIN-SEINUNIT

Table 3.6: EHF_ZZ_UHF: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **active_beg_qtr_es** | First QTIME with positive employment | Num | 3 |
| **active_employ_es** | = ...1.. if positive employment in QTIME i | Char | 124 |
| **active_end_qtr_es** | Last QTIME with positive employment | Num | 3 |
| **active_ever_es** | Ever had positive employment | Num | 8 |
| **active_qtrs_es** | Number of quarters with positive employment | Num | 3 |
| **bpemp_es1** | Month 1 employment in QTIME=1 | Num | 8 |
| **...** | | | |
| **bpemp_es124** | Month 1 employment in QTIME=124 | Num | 8 |
| **emp_es1** | Maximum monthly employment in QTIME=1 | Num | 8 |
| **...** | | | |
| **emp_es124** | Maximum monthly employment in QTIME=124 | Num | 8 |
| **mu_code** | ...1... if part of multi-establishment, ...2... if master unit | Char | 124 |
| **numruns1** | Number of establishments in QTIME=1 | Num | 8 |
| **...** | | | |
| **numruns124** | Number of establishments in QTIME=124 | Num | 8 |
| **sein** | Standardized State Employer ID Number | Char | 12 |
| **seinunit** | State UI reporting unit | Char | 5 |

#### 3.4.4.2 SHF

The SHF (SEIN History File) contains a full employment history at the firm level, for each SEIN (wide file).

**Record identifier** SEIN
**Sort order** SEIN
**Entity** Firm
**Unique Entity Key** SEIN

Table 3.7: EHF_ZZ_SHF: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **active_beg_qtr_es** | First QTIME with positive employment | Num | 3 |
| **active_employ_es** | = ...1.. if positive employment in QTIME i | Char | 124 |
| **active_end_qtr_es** | Last QTIME with positive employment | Num | 3 |
| **active_ever_es** | Ever had positive employment | Num | 8 |
| **active_qtrs_es** | Number of quarters with positive employment | Num | 3 |
| **bpemp_es1** | Month 1 employment in QTIME=1 | Num | 8 |
| **...** | | | |
| **bpemp_es124** | Month 1 employment in QTIME=124 | Num | 8 |
| **emp_es1** | Maximum monthly employment in QTIME=1 | Num | 8 |
| **...** | | | |
| **emp_es124** | Maximum monthly employment in QTIME=124 | Num | 8 |
| **estabs_es1** | in QTIME=1 | Num | 8 |
| **...** | | | |
| **estabs_es124** | in QTIME=124 | Num | 8 |
| **ever_mu** | SEIN has ever had multiple units | Num | 8 |
| **sein** | Standardized State Employer ID Number | Char | 12 |

### 3.4.5 Files with cross-state information

The following files contain cross-state information. Access to these files does not depend on state-specific permission, since no information in the files itself is provided by the states.

#### 3.4.5.1 National Indicator File

The EHF National Indicator File provides information to researchers on the presence of wage records for workers in other states, even when the researcher does not permissions to access those other states' data files. It tabulates, in a format similar to the core EHF files, the number of states in which a worker was observed to have jobs. It does not provide the total number of jobs, nor the earnings in those other states. It is meant to be used as a control in situations where sample bias might be an issue. However, users should note that only at-risk states are counted, see 3.4.5.2.

**Record identifier** PIK YEAR<br>
**Sort order** PIK YEAR<br>
**Entity** Person<br>
**Unique Entity Key** PIK

Table 3.8: EHF_US_INDICATORS: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **num_states_earn1** | Number of states with positive earnings, Q1 | Num | 8 |
| **num_states_earn2** | Number of states with positive earnings, Q2 | Num | 8 |
| **num_states_earn3** | Number of states with positive earnings, Q3 | Num | 8 |
| **num_states_earn4** | Number of states with positive earnings, Q4 | Num | 8 |
| **pik** | Protected Identification Key | Char | 9 |
| **year** | Calendar year | Num | 3 |

#### 3.4.5.2 States-at-risk File

The EHF States-at-risk File provides information to researchers on the availability of wage records for each state, with start and end dates (year and quarter) identifying when states were at risk for workers to appear in them. In combination with the National Indicator File (3.4.5.1), it is meant to be used as a control in situations where sample bias might be an issue.

**Record identifier** state
**Sort order**
**Entity** state
**Unique Entity Key** state

Table 3.9: EHF_ALL_AVAILABILITY: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **end_quarter** | End quarter of the data within the table | Num | 8 |
| **end_year** | End year of data within the table | Num | 8 |
| **start_quarter** | Start quarter of the data within the table | Num | 8 |
| **start_year** | Start year of data within the table | Num | 8 |
| **state** | 2-digit State Code | Char | 2 |

## 3.5 HELPFUL PROGRAMS

The following programs might be found to be useful when using the data.

### 3.5.1 Selecting a random subsample of persons

The following program allows users to select a random sample of approximately one percent of individuals. It relies on the fact that the first two characters of the PIK are approximately uniformly distributed on $[00, 99]$. Note that 'AA' is a valid value for the first two characters and denotes individuals for whom no valid SSN was on file. Occurrence of such "pseudo-PIKs" varies by state.

This file needs to be run across all available states, and then unduplicated before merging against a raw analysis file.

```
%let syear=2014;

data my_piklist;
set INLIB.ehf_ca_uniqpik(where=(cut='01'));
by pik;
run;

data my_analysis;
    merge my_piklist(keep=pik in=a)
          my_rawdata(keep=analysis variables pik)
          ;
    by pik;
    if a;
 run;
```

## 3.6 NOTES

A note of caution: this list may be incomplete.

Table 3.10: UI/EHF Summary of Information and Known Issues with Data Coverage and Quality

| State | Known Data Quality Issues (UI/EHF) | Recommendation to Researchers |
|---|---|---|
| CA | None | |
| CO | 60-70% hole in UI data in 1993:3. 20% unresolved identifier mismatch on UI in [90:1-90:3] | Researchers should generally avoid use of pre-1994 EHF data in CO. |
| FL | (1) There appear to be changes being made in the firm identifiers on the ES202 and UI data in the mid-to-late 1990s. Specifically it looks as though some changes are made on the identifiers in the ES202 in 1996 and in 1997 the UI data is corrected in kind. In the ES202 data, 14% of firms die in 1995:4 and are born in 1996:1, indicating a shift in some firm identifiers. A similar change in magnitude occurs in the UI data between 1997:1 and 1997:4. Between these years, the rate of match between the UI and ES202 SEINs is somewhat poor ( 10% of UI SEINs do not appear on the ES202 between 1996:1 and 1997:3), although it is quite good both before and after. (2) The match between the ES202 and UI data is not good in 2002:4-2003:3, with 13-20% of UI SEINs not appearing in the ES-202 data. | While not a big enough problem to recommend avoiding use of these date ranges in FL, be aware that changes in firm identifiers in the mid-1990s will bias worker flow measures during this period. |
| IA | None | |
| ID | 1990 UI data has firm identifier problems on approximately 40% of the records. Because of these problems, this year is not included in the EHF. | Researchers should generally avoid use of 1990 ID EHF data, which should not be too much of an issue as ES202 information is missing for this year in ID. |
| IL | Small hole in UI data in 1990:1 ( 10% missing). 1992:1 and 1993:1 are also missing UI wage records. | Note to researchers: These problems bias worker flows in those quarters, also full quarter employment in early years of IL data. |
| IN | None | |
| KS | Large holes in KS UI data at 1990:1 (>50% missing) and 1992:4 (25% missing) | Researchers should generally avoid use of 1990 and 1992 KS EHF data; this problem will also bias full quarter employment and flows in 1993. |
| KY | UI identifier problem in 2000:3-2001:2 likely, due to 10%, 15% death rates in 2000:3, 2000:4, followed by 11%, 14% birth rates in 2001:1 and 2001:2. (Normal is 3-7% births/deaths in a particular quarter) | Note to researchers: These problems bias worker flows in those quarters, also full quarter employment during 2000-2001 KY data. |
| MD | None | |
| ME | None | |

(cont)

Table 3.10 – Continued

| State | Known Data Quality Issues (UI/EHF) | Recommendation to Researchers |
|---|---|---|
| MN | None | |
| MO | 1994:4 UI data is small (approximately 70% sample). | Researchers should generally avoid use of 1994 MO EHF data; this problem will also bias some full quarter employment and flows measures in 1995. |
| MT | | |
| NC | * ES202 show persistently lower employment than UI, by about 14%, except for 1991:1-1992:3 (around 0%) and 2002:1-2002:4 (5-8%). Warnings are generated when it goes above 15%. * Payroll is typically 6-8% higher on ES202 compared to UI except for 1991:1-1992:3, where it is 20-30% higher. There are also significant, but not as large deviations in 2002:1-2003:1.<br>* Based on the BLS PU records, the ES202 data series looks fine: ES202 sums rarely go above 1% (Test 13-1 and 13-2)<br>Conclusion: we are still missing wage records in the early periods, and some in later periods as well. The most recent wage records actually look coherent with the longest time series, but 2002 is a small problem. | Note to Researchers: Similar to problems in early years of IL, these issues bias worker flows in those quarters, also full quarter employment. |
| NJ | Small holes in NJ UI data at 1998:3 ( 5%) and 1999:1 ( 8-10%) and 2003:1 ( 10%) | Note to Researchers: Problem probably small enough to ignore for most research purposes. |
| NM | None | |
| OK | None | |
| OR | 1994:1 is small, but not terribly so. | Note to Researchers: Problem probably small enough to ignore for most research purposes. |
| PA | UI wage records are 1% sample for 1996:4 | Note to Researchers: Generally avoid use of 1996 PA annual earnings (particularly earnings changes between 1995-1996, 1996-1997, which will be biased), this problem will also bias some flows and full quarter employment measures in 1996 and 1997. |
| TX | None | |
| VA | 1998:1 is small, and 1998:2 also looks on the small side. | Note to Researchers: Problems probably small enough to ignore for most research purposes. |
| WA | None | |
| WI | None | |
| WV | None | |

# Chapter 4.
# Geocoded Address List (GAL)

## 4.1 OVERVIEW

The Geocoded Address List (GAL) core dataset contains a state's unique commercial and residential addresses with geocodes to indicate Census geography (state/county/tract/block) and geographic coordinates (latitude/longitude). For each state, the GAL consists of an address list (GAL core), a crosswalk for each address input source, and other ancillary datasets. The GAL core contains each standardized, deduplicated, unique address; a unique identifier called `galid`; geocodes; and processing information. The GAL crosswalks link GAL core `galids` and specific input source addresses; each input address has a unique identifier called `bigsrcid`.

## 4.2 DATA CITATION

> U.S. Census Bureau. 2016. *Geo-coded Address List (GAL) in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

### 4.2.1 Changes in this Snapshot

The GAL is under continuous development and improvement. In this snapshot, the GAL has been completely restructured. Table 4.1 provides an overview of the available files. Due to fundamental changes to the structure of the GAL, an attempt was made to maintain backward compatibility, but not all changes could be reasonably mapped. Crosswalks to all input sources have been re-introduced. Access to certain crosswalks may require additional permissions or may not be available to RDC-based researchers due to legal or contractual restrictions.

Table 4.1: GAL components

| Dataset | Description |
|---|---|
| gal_zz_core | GAL core: deduplicated addresses with geographic coordinates, Census geography, indicator of contributing source(s), and quality indicators |
| gal_zz_core_t26 | Ancillary GAL core containing records sourced solely from SSEL and Business Register (**FTI**) |
| gal_zz_core_es202_only | Ancillary GAL core containing records sourced solely from ES202 files (**available to internal Census projects only**) |
| | Crosswalk between input addresses (all years) and GAL core addresses for each input source: |
| gal_zz_xwalk | ES202 (**available to internal Census projects only**) |
| gal_zz_acspow | American Community Survey - Place of Work |
| gal_zz_ahs | American Housing Survey |
| gal_zz_maf | Master Address File (MAF) |
| gal_zz_nbr_t26 | Business Register (**FTI**) |
| gal_zz_ssel_t26 | Standard Statistical Establishment List (SSEL) (**FTI**) |
| gal_zz_YYYY_bmf | Block Map File (BMF) with the geovintage indicated in the filename; contains the various geographies that each block belongs to: WIB area, county, core-based statistical area, place, etc. |
| gal_zz_YYYY_tccb | TIGER County Centroid Blocks (TCCB) with the geovintage indicated in the filename; contains the BMF values for one block near each county's geographic center; considered the default latitude/longitude and geographies for each county |

## 4.3 DETAILED DESCRIPTION

### 4.3.1 Input Data

The input data consists of addresses, geocodes, and coordinates. As of early 2013, the source files providing addresses consisted of the following:

| | |
|---|---|
| ES202 | QCEW (all available years 1990 and later) |
| SSEL | Standard Statistical Establishment List (1990 - 2001) |
| BR | Business Register (2002-2010) |
| MAF | Master Address File (2006-2014) |
| ACS-POW | American Community Survey Place of Work file (2001 through 2007) |
| AHS | American Housing Survey (2002) |

### 4.3.2 Geocodes

The following sources provide Census geography and geographic coordinates:

| | |
|---|---|
| Spectrum | Pitney Bowes Spectrum Technology Platform, a geocoding software |
| MAF | Master Address File, from U.S. Census Bureau, Geography Division |
| GRF-C | Geographic Reference File - Codes (encompassed in the BMF), from U.S. Census Bureau, Geography Division |
| WIB-C | Workforce Investment Board - Codes (encompassed in the BMF) |
| BMF | Block Map File |

### 4.3.3 Processing description

The following job stream describes GAL processing at a high level. Each step uses SAS, unless otherwise indicated. The GAL is incremental; the datasets are maintained quarter to quarter and only new input sources are added during regular production. The numbers presented do not align with GAL process module numbers.

1. Identify newly available input sources and read in their addresses.
2. Prepare new addresses to run through the address standardizer/geocoder. Prepare previous quarters' addresses for a subsequent run if they have not yet been geocoded to their expected quality level.
3. Assign "expert geocodes" to input addresses that match records in a special look-up table. On a limited basis, some poor geocoding outcomes have triggered research and we have assigned specific geocodes to specific input addresses or establishments in a look-up table. The GAL assigns those geocodes every time an input address or establishment matches a record in that table.
4. Run the addresses prepared in step 2 through Spectrum's address standardization and geocoding modules, using a script. The standardization module makes minor corrections to addresses and standardizes components like directional and street types (e.g., '`123 North Main Street`' might standardize to '`123 N Main St`').[1] The geocoding module assigns Census geography (state/county/tract/block) and geographic coordinates to each address, when possible.
5. Match standardized addresses from step 4 to other new and existing standardized addresses, based on an exact match of the address (e.g., '`123 N Main St`' to '`123 N Main St`').

[1] All addresses in this section are fake and made up to illustrate specific points.

6. Parse any unmatched addresses from step 5 using SAS Data Quality and attempt another match. For example, ‘`123 N Main St Unit 4`’ is parsed into house number ‘`123`’, pre-directional ‘`N`’, street ‘`Main`’, street type ‘`St`’, and extension number ‘`4`’. This allows it to match to the existing GAL address ‘`123 N Main St Ste 4`’.
7. Create preliminary datasets and crosswalks from the matched-address clusters.
8. Assign to each address the best geocodes from each address cluster’s various candidates (MAF geocodes are preferred over others).
9. For addresses missing geocodes, attempt to derive the geocodes from the MAF. For example, the MAF might contain geocodes for ‘`121 Main St`’ and ‘`125 Main St`’. If the GAL’s ‘`123 Main St`’ is missing geocodes, the GAL can reasonably assume that ‘`123`’ exists between the MAF’s ‘`121`’ and ‘`125`’; it will assign to ‘`123`’ the same Census geography as its neighbors. The GAL can also interpolate geographic coordinates.
10. For addresses with tract-level geocodes (i.e., missing a block), impute a block based on the distribution of similar addresses within the tract (e.g., other commercial addresses or other residential addresses).
11. Output finalized datasets and crosswalks.

### 4.3.4 Description of RDC files

The GAL core dataset (Section 4.4.3) containing standardized and deduplicated addresses is available in the RDC environment, though some data are suppressed. Crosswalks to the American Community Survey Place of Work file (ACS-POW) (4.4.6.1), American Housing Survey (AHS) (4.4.6.2), and Master Address File (MAF) (4.4.6.3) input sources are also available. Crosswalks to FTI input sources are BR (4.4.6.4) and SSEL (pre-2002 Business Register, 4.4.6.5).

The Census-internal GAL is commingled data: it contains information protected under both Title 13 and Title 26. Before transferring the GAL to the RDC environment, the data are split into components that are Title 26-protected, and those that are not. In particular, we remove any record sourced solely from Title 26 data (BR or SSEL) from the GAL core dataset and store the information separately in `gal_zz_core_t26.sas7bdat` (Section 4.4.4).

Due to restrictions embodied in LEHD’s MOU with its partner states, address data sourced in the ES-202 (QCEW) files cannot be made available to non-Census researchers. Variables containing address information sourced solely from ES-202 are blanked in `gal_zz_core`, but are preserved in `gal_zz_core_es202_only`.

Geocodes, i.e., geographic coordinates and Census geography, remain intact in the core dataset. If a record is sourced from both a Title 26 input source and ES-202, but not from other input sources, then we include the complete record in `gal_zz_core_t26`, and a truncated record (without address data) in `gal_zz_core`. In these instances, the same address record (with identical galid) will exist in both the `gal_zz_core` and `gal_zz_core_t26` datasets, though the address variables will be empty in the GAL core dataset and populated in the Title 26 dataset.

Note that crosswalks are of no use without the file that they point to. A researcher must request the input files separately. Not all input files described here may be available in the RDC environment in the form that they are used by LEHD.

### 4.3.5 Important Variables

**Unique identifier**

The GAL’s unique address identifier variable is `galid`, a 29-character string. Beginning in the fourth quarter of 2013, an address’s `galid` is consistent between GAL vintages.

| Structure of galid | |
|---|---|
| Position 1 | the letter 'A' |
| Position 2-14 | vintage identifier indicating when the GAL initially added the address |
| Position 15 | underscore character |
| Position 16-17 | state FIPS code |
| Position 18-29 | zero-padded sequential number |

### Geographic codes

The variable geoid_2010 identifies the Census geography assigned to each address.

| | |
|---|---|
| Position 1-2 | state FIPS code |
| Position 3-5 | county FIPS code |
| Position 6-11 | Census tract |
| Position 12-15 | Census block (position 12 indicates the Census block group) |

The variable a_galgeoyr in gal_zz_core indicates the vintage for the Census geography found in the geoid_2010 variable and is consistent throughout the GAL. The GAL consistently uses Decennial Census tabulation geography for an entire decade (currently 2010 geography).

The higher-level geographies assigned to a block can change in the years between Decennial Censuses. For example, a county boundary change will result in some blocks being assigned to a different county. A city's annexation will result in some blocks being assigned to a different place. The BMF (Section 4.4.7.1) contains higher-level geographies that are current for the year indicated in the BMF's filename. For example, an address assigned the geoid_2010 of **515150501005036** would indicate that the state/county FIPS code was **51515** in 2010. That block had since been absorbed by another county; the 2014 BMF's a_st and a_cty variables indicate that the current state/county FIPS code for that block is **51019**.

The BMF's unique key is geoid_2010. A researcher can match-merge or join gal_zz_core (and/or gal_zz_core_t26, if accessible) and any available BMF vintage to determine GAL core addresses' higher-level geographies for a desired year.

To summarize: the geoid_2010 variable will consistently represent 2010 Census geography. A researcher can use that variable to obtain current geographies from the BMF (current as of the year indicated in the filename).

The variable a_block_src indicates the source of an address's assigned block:

| Value | Typical Percent | Description |
|---|---:|---|
| M | 79.36 | MAF |
| C | 9.83 | Pitney Bowes Spectrum (commercial geocoder) |
| X | 0 | Expert geocode (clerical research) |
| O | 4.31 | Tract known, block imputed using commercial addresses |
| I | 0.08 | Tract known, block imputed using mixed addresses |
| S | 3.3 | Tract known, block imputed using residential addresses |
| D | 0.01 | Tract known, block imputed using all addresses |
| E | 0.01 | Derived from MAF (street number equal to a neighbor) |
| W | 0.08 | Derived from MAF (between two neighbors) |
| N | 0.25 | Out-of-state GAL record (geocode is set to blank) |
| missing | 2.78 | Missing a_block_src code |

### Geographic Coordinates

The variables `a_latitude` and `a_longitude` indicate each address's geographic coordinates. These variables are numeric with six implied decimals (divide by 1,000,000 to convert them); however, the coordinates are not as precise as six decimal places imply.

The numeric variable `a_geoqual` is a ranked quality indicator of the coordinates' precision, where 1 is best and 9 is worst. It can be interpreted roughly as precision relative to a level of Census geography:

| Value | Typical Percent | Interpretation |
|---|---:|---|
| 1 (best) | 85.38 | Block (Rooftop accuracy) |
| 2 | 3.19 | Block (ZIP+4) |
| 3 | 0.01 | Block group |
| 4 | 8.4 | Tract |
| 9 (worst) | 3.02 | County |

The `a_latlong_src` variable indicates the source of an address's assigned coordinates:

| Value | Typical Percent | Description |
|---|---:|---|
| M | 64.62 | MAF |
| C | 23.85 | Pitney Bowes Spectrum (commercial geocoder) |
| B | 8.41 | Block internal point |
| D | 0.1 | Derived |
| X | 0 | Expert geocode (clerical research) |
| missing | 3.02 | Coordinates are missing |

Few addresses have a_latlong_src equal to ’D’. Coordinates derivation occurs only if coordinates are still missing after Pitney Bowes Spectrum processing and direct extraction from the MAF, but the tract is known. In this case, the flag a_latlong_drv describes the derivation method:

| Value | Typical Percent | Description |
|---|---:|---|
| F | 0.01 | Adopted from the only address on the block face |
| P | 0.1 | Extrapolated between 2 addresses on the block face |
| missing | 99.89 | Derivation not performed |

**Source ID**

The bigsrcid variable in the crosswalk datasets is a unique key for every address from every input source used in the GAL.

The bigsrcid itself is constructed in a way that identifies the input source, year (and quarter, if ES-202), and record for each input address. The bigsrcid construction layout varies by input source, as shown below:

- In position 31, ’P’ or ’S’ indicates Primary or Secondary. Currently, all addresses are Primary.
- In position 32, ’P’ or ’M’ indicates Physical or Mailing.
- In position 33, ’R’ or ’C’ indicates Residential or Commercial.

**ACS-POW**

| 1 | 2 - 5 | 6 - 14 | 15 | 16 - 17 | 18 - 19 | 20 - 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|---|---|---|
| ’P’ | year | cmid | seq | pnum | acsfileseq | ’00000000000’ | ’P’ or ’S’ | ’P’ or ’M’ | ’R’ or ’C’ |

**AHS**

| 1 | 2 - 5 | 6 - 18 | 19 - 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|
| ’H’ | year | control | ’000000000000’ | ’P’ or ’S’ | ’P’ or ’M’ | ’R’ or ’C’ |

**Business Register**

| 1 | 2 - 5 | 6 - 15 | 16 - 24 | 25 | 26 - 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|---|---|
| ’N’ | year | empunit_id_char | ein | singmult | ’00000’ | ’P’ or ’S’ | ’P’ or ’M’ | ’R’ or ’C’ |

**ES-202**

| 1 | 2 - 5 | 6 - 17 | 18 - 22 | 23 - 26 | 27 | 28 - 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|---|---|---|
| ’E’ | year | sein | seinunit | ’0000’ | quarter | ’000’ | ’P’ or ’S’ | ’P’ or ’M’ | ’R’ or ’C’ |

**MAF**

| 1 | 2 - 5 | 6 - 18 | 19 - 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|
| ’M’ | year | mafid | ’000000000000’ | ’P’ or ’S’ | ’P’ or ’M’ | ’R’ or ’C’ |

**SSEL**

| 1 | 2 - 5 | 6 - 15 | 16 | 17 - 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|---|
| ’B’ | year | cfn | singmult | ’00000000000000’ | ’P’ or ’S’ | ’P’ or ’M’ | ’R’ or ’C’ |

### 4.3.6 Accessing the GAL: the GAL Crosswalks

The GAL crosswalks link GAL core addresses and input addresses, as each crosswalk record contains a galid and bigsrcid (as well as the bigsrcid’s components). This allows you to extract geocodes and standardized addresses from the GAL for an entity (e.g., an ES-202 establishment or MAF housing unit). The entity ID variables required to identify an entity vary based on the input source:

| | |
|---|---|
| ES-202 (gal_zz_xwalk) | sein, seinunit, year, quarter |
| Business Register (gal_zz_nbr_t26) | empunit_id_char, year, singmult |
| SSEL (gal_zz_ssel_t26) | cfn, singmult, year |
| MAF (gal_zz_maf) | mafid, year |
| ACS-POW (gal_zz_acspow) | cmid, seq, pnum, acsfileseq, year |
| AHS (gal_zz_ahs) | control, year |

The crosswalks also contain flag variables. A single entity can be associated with multiple GAL core address records. For example, an ES-202 establishment might provide both physical and mailing addresses. Furthermore, one input address might match multiple GAL core addresses. The input address "123 Main St" might match both "123 N Main St" and "123 S Main St". The flags are:

isbest: Indicates that the GAL core address (galid) is the best candidate for an entity.
iseq: Sequential number (but not a ranking) assigned to each candidate returned by Spectrum for a single input address.
e_flag, n_flag, b_flag, m_flag, p_flag, h_flag: For ES-202, Business Register, SSEL, MAF, ACS-POW, and AHS, respectively; indicates whether the input address is a physical or mailing address.
prisec: Indicates whether the address is primary or secondary (currently all are primary).

To determine all contributing input sources for a specific GAL core address record or records, find the specific galid(s) in all available crosswalks: gal_zz_xwalk, gal_zz_maf, gal_zz_acspow, gal_zz_ahs; plus gal_zz_nbr_t26 and gal_zz_ssel_t26, if accessible.

## 4.4 DATA SET DESCRIPTIONS

### 4.4.1 Naming scheme

GAL files are labeled with the geovintage used in the creation, when relevant. By design, multiple geovintages are available for the same data point, allowing to crosswalk between geovintages.

SAS datasets with zero observations are attached to this document:[2]

- galcc/zz/gal_zz_core_es202only.sas7bdat
- galt26/zz/gal_zz_core_t26.sas7bdat
- galt26/zz/gal_zz_nbr_t26.sas7bdat
- galt26/zz/gal_zz_ssel_t26.sas7bdat
- gal/zz/gal_zz_2010_bmf.sas7bdat
- gal/zz/gal_zz_2010_tccb.sas7bdat
- gal/zz/gal_zz_2012_bmf.sas7bdat
- gal/zz/gal_zz_2012_tccb.sas7bdat
- gal/zz/gal_zz_2013_bmf.sas7bdat
- gal/zz/gal_zz_2013_tccb.sas7bdat
- gal/zz/gal_zz_2014_bmf.sas7bdat
- gal/zz/gal_zz_2014_tccb.sas7bdat
- gal/zz/gal_zz_acspow.sas7bdat
- gal/zz/gal_zz_ahs.sas7bdat
- gal/zz/gal_zz_core.sas7bdat
- gal/zz/gal_zz_maf.sas7bdat
- gal/zz/gal_zz_xwalk.sas7bdat

`ZZ` stands for the state postal abbreviation, and `YYYY` for the geovintage year. Not all files are available for all states. In particular, LEHD-related crosswalks are only available for states actively participating with LEHD at the time of creation of the GAL.

### 4.4.2 Data location

The files are stored in two main directories, with state-specific subdirectories:

```
gal/zz/       for most files
galt26/zz     for files with Title 26 protected content
```

[2] Also visible on the attachment tab - Adobe Reader may be required.

### 4.4.3 Main dataset: GAL_ZZ_core

This file does not contain data protected exclusively under Title 26; see Section 4.4.4. This file also does not report any address data sourced exclusively from ES-202; see Section 4.4.5. If a record contains fields sourced exclusively from FTI data or from ES-202 data (address data), the values have been blanked on this file.

**Record identifier:** GALID
**Sort order:** GALID
**File indexes:** none
**Entity** unique address
**Unique Entity Key** GALID

Table 4.2: gal_zz_core: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **a_block_src** | Block assignment source: assigned by geoCoder or Maf, derived from maf, Equals or betWeen maf addresses, or imputed from addresses in tract: cOmmercial, reSidential, mIxed, or all aDdre | Char | 1 |
| **a_galgeoyr** | Decennial Census from which GAL geocodes are based | Num | 8 |
| **a_geoqual** | Quality of lat/long | Num | 3 |
| **a_geoqual_issue** | Inconsistency between a_geocode and a_geoqual | Num | 8 |
| **a_geoqual_original** | Original a_geoqual value assigned by geocoder, GAL can override this under certain conditions | Num | 8 |
| **a_latitude** | Latitude, 6 implied decimal places | Num | 8 |
| **a_latlong_drv** | Coordinates derivation method: only address on blockFace or extraPolation | Char | 1 |
| **a_latlong_src** | Coordinates assignment source: Maf, geoCoder, Derived, Block internal point | Char | 1 |
| **a_longitude** | Longitude, 6 implied decimal places | Num | 8 |
| **galid** | Unique GAL address ID: 'A' \|\| GAL vintage when added \|\| sequential integer | Char | 29 |
| **gc_address** | Address with unit | Char | 80 |
| **gc_city** | Address city | Char | 35 |
| **gc_country** | Address country | Char | 3 |
| **gc_path** | During last GAL vintage, G = address was run through geocoder and P = Passed with existing geocodes | Char | 1 |
| **gc_state** | Address state abbreviation | Char | 2 |
| **gc_vintdate** | Geocoder vintage | Char | 4 |
| **gc_zip** | ZIP Code | Char | 5 |
| **gc_zip4** | ZIP+4 Code | Char | 4 |
| **geoid_2010** | Census geocode: Tabulation state FIPS (2) \|\| Tabulation county FIPS (3) \|\| tract (6) \|\| block (4) | Char | 15 |
| **m_block_src** | e = MAF block is known but not used | Char | 1 |

### 4.4.4 Auxiliary dataset: GAL_ZZ_core_T26

This file has the same column structure as the main file, but contains all records sourced exclusively from Title 26-protected information. The columns are described in Section 4.4.3.

### 4.4.5 Auxiliary dataset: GAL_ZZ_core_es202only

This file has the same column structure as the main file, but contains all records sourced exclusively from QCEW-sourced information. These data are only available to internal Census projects. The columns are described in Section 4.4.3.

### 4.4.6 Crosswalks

#### 4.4.6.1 ACS-POW crosswalk: GAL_ZZ_acspow

**Record identifier:** cmid seq pnum acsfileseq year iseq
**Sort order:** cmid seq pnum acsfileseq year iseq
**File indexes:** none
**Entity** Address of an ACS respondent in a particular year
**Unique Entity Key** cmid seq pnum acsfileseq year

Table 4.3: GAL_ZZ_ACSPOW: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **acsfileseq** | ACS file sequence number | Char | 2 |
| **addr_qtime** | Quarter index (1985Q1=1) | Num | 3 |
| **bigsrcid** | Unique input address ID | Char | 36 |
| **cmid** | Continuous measurement ID | Char | 9 |
| **galid** | Unique GAL address ID: 'A' \|\| GAL vintage when added \|\| sequential integer | Char | 29 |
| **isbest** | Best GAL address candidate for entity/quarter | Char | 1 |
| **iseq** | Unranked identifier for address candidate returned from geocoder for a single input address | Char | 3 |
| **p_flag** | Physical or Mailing | Char | 1 |
| **pnum** | Person number | Char | 2 |
| **prisec** | Primary or Secondary | Char | 1 |
| **seq** | Sequence number | Char | 1 |
| **year** | Year YYYY | Char | 4 |

#### 4.4.6.2 AHS crosswalk: GAL_ZZ_ahs

**Record identifier:** control year iseq
**Sort order:** control year iseq
**File indexes:** none
**Entity** Address of a AHS household in a particular year
**Unique Entity Key** control year

Table 4.4: GAL_ZZ_AHS: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **addr_qtime** | Quarter index (1985Q1=1) | Num | 3 |
| **bigsrcid** | Unique input address ID | Char | 36 |
| **control** | Control | Char | 13 |
| **galid** | Unique GAL address ID: 'A' \|\| GAL vintage when added \|\| sequential integer | Char | 29 |
| **h_flag** | Physical or Mailing | Char | 1 |
| **isbest** | Best GAL address candidate for entity/quarter | Char | 1 |
| **iseq** | Unranked identifier for address candidate returned from geocoder for a single input address | Char | 3 |
| **prisec** | Primary or Secondary | Char | 1 |
| **year** | Year YYYY | Char | 4 |

#### 4.4.6.3 MAF crosswalk: GAL_ZZ_maf

**Record identifier:** mafid year iseq
**Sort order:** mafid year iseq
**File indexes:** none
**Entity** Address on MAF
**Unique Entity Key** mafid year

Table 4.5: GAL_ZZ_MAF: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **addr_qtime** | Quarter index (1985Q1=1) | Num | 3 |
| **bigsrcid** | Unique input address ID | Char | 36 |
| **galid** | Unique GAL address ID: ’A’ \|\| GAL vintage when added \|\| sequential integer | Char | 29 |
| **isbest** | Best GAL address candidate for entity/quarter | Char | 1 |
| **iseq** | Unranked identifier for address candidate returned from geocoder for a single input address | Char | 3 |
| **m_flag** | Physical or Mailing | Char | 1 |
| **mafid** | Master Address File ID | Char | 12 |
| **prisec** | Primary or Secondary | Char | 1 |
| **year** | Year YYYY | Char | 4 |

#### 4.4.6.4 Business Register crosswalk: GAL_ZZ_nbr_t26

Crosswalk to Business Register (BR) (2002-2010). For more information on the BR, see DeSalvo, Limehouse, and Klimek (2016). The crosswalk itself is FTI, and requires permission from Internal Revenue Service (IRS) to use.

**Record identifier:** empunit_id_char year singmult iseq
**Sort order:** empunit_id_char year singmult iseq
**File indexes:** none
**Entity** Establishment in a particular year
**Unique Entity Key** empunit_id_char singmult year

Table 4.6: GAL_ZZ_NBR_T26: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **addr_qtime** | Quarter index (1985Q1=1) | Num | 3 |
| **bigsrcid** | Unique input address ID | Char | 36 |
| **empunit_id_char** | Business Register employer unit number | Char | 10 |
| **galid** | Unique GAL address ID: 'A' \|\| GAL vintage when added \|\| sequential integer | Char | 29 |
| **isbest** | Best GAL address candidate for establishment/quarter | Char | 1 |
| **iseq** | Unranked identifier for address candidate returned from geocoder for a single input address | Char | 3 |
| **n_flag** | Physical or Mailing | Char | 1 |
| **prisec** | Primary or Secondary | Char | 1 |
| **singmult** | Single-unit or Multi-unit | Char | 1 |
| **year** | Year YYYY | Char | 4 |

#### 4.4.6.5 SSEL crosswalk: GAL_ZZ_ssel_t26

Crosswalk to Standard Statistical Establishment List (SSEL), predecessor to the BR (1990 through 2001). For more information on the SSEL and the BR, see DeSalvo, Limehouse, and Klimek (2016). The crosswalk itself is FTI, and requires permission from IRS to use.

**Record identifier:** cfn singmult year iseq
**Sort order:** cfn singmult year iseq
**File indexes:** none
**Entity** Establishment in a particular year
**Unique Entity Key** cfn singmult year

Table 4.7: GAL_ZZ_SSEL_T26: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **addr_qtime** | Quarter index (1985Q1=1) | Num | 3 |
| **b_flag** | Physical or Mailing | Char | 1 |
| **bigsrcid** | Unique input address ID | Char | 36 |
| **cfn** | Census File Number | Char | 10 |
| **galid** | Unique GAL address ID: ’A’ \|\| GAL vintage when added \|\| sequential integer | Char | 29 |
| **isbest** | Best GAL address candidate for establishment/quarter | Char | 1 |
| **iseq** | Unranked identifier for address candidate returned from geocoder for a single input address | Char | 3 |
| **prisec** | Primary or Secondary | Char | 1 |
| **singmult** | Single-unit or Multi-unit | Char | 1 |
| **year** | Year YYYY | Char | 4 |

## 4.4.7 Auxiliary data

### 4.4.7.1 Block Map File: GAL_ZZ_YYYY_bmf

Block Map File (BMF) contains higher-level geographies for each block, extracted from the Geographic Reference File - Codes (GRF-C). YYYY indicates the geovintage.

**Record identifier:** geoid_2010
**Sort order:** geoid_2010
**File indexes:** none
**Entity** Census block in 2010
**Unique Entity Key** geoid_2010

Table 4.8: GAL_ZZ_2010_BMF: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **a_block** | Current census block | Char | 4 |
| **a_block_suf1** | Current census block suffix 1 | Char | 1 |
| **a_block_suf2** | Current census block suffix 2 | Char | 1 |
| **a_cbsa** | Core-Based Statistical Area (current) | Char | 5 |
| **a_cbsa_memi** | CBSA type: 1=Metro, 2=Micro, 9=not in CBSA | Char | 1 |
| **a_cty** | Current county FIPS | Char | 3 |
| **a_cty_tab** | Tabulation county FIPS | Char | 3 |
| **a_fipsmcd** | Minor Civil Division FIPS (current) | Char | 5 |
| **a_fipspl** | Place FIPS (current) | Char | 5 |
| **a_geocode** | Tabulation state FIPS \|\| Tabulation county FIPS \|\| tabulation census tract | Char | 11 |
| **a_polylat** | Block's internal point latitude, 6 implied decimals | Num | 8 |
| **a_polylong** | Block's internal point longitude, 6 implied decimals | Num | 8 |
| **a_ssccc** | Current state-county FIPS | Char | 5 |
| **a_st** | Current state FIPS | Char | 2 |
| **a_st_tab** | Tabulation state FIPS | Char | 2 |
| **a_taz** | Traffic Analysis Zone (current) | Char | 8 |
| **a_tract** | Current census tract | Char | 6 |
| **a_wib** | Workforce Investment Board area (current) | Char | 6 |
| **geoid_2010** | Census geocode: Tabulation state FIPS (2) \|\| Tabulation county FIPS (3) \|\| tract (6) \|\| block (4) | Char | 15 |
| **geovtg** | Geo-vintage for current geography | Num | 8 |

### 4.4.7.2 TCCB: GAL_ZZ_YYYY_tccb

TIGER County Centroid Blocks (TCCB) with the geovintage indicated in the filename; contains the BMF values for one block near each county's geographic center; considered the default latitude/longitude and geographies for each county. YYYY indicates the geovintage.

**Record identifier:** geoid_2010
**Sort order:** geoid_2010
**File indexes:** none
**Entity** Census block in 2010
**Unique Entity Key** geoid_2010

Table 4.9: GAL_ZZ_2010_TCCB: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **a_block** | Current census block | Char | 4 |
| **a_block_src** | Block assignment source: A=ArcGIS processing | Char | 1 |
| **a_block_suf1** | Current census block suffix 1 | Char | 1 |
| **a_block_suf2** | Current census block suffix 2 | Char | 1 |
| **a_cbsa** | Core-Based Statistical Area (current) | Char | 5 |
| **a_cbsa_memi** | CBSA type: 1=Metro, 2=Micro, 9=not in CBSA | Char | 1 |
| **a_cty** | Current county FIPS | Char | 3 |
| **a_cty_tab** | Tabulation county FIPS | Char | 3 |
| **a_fipsmcd** | Minor Civil Division FIPS (current) | Char | 5 |
| **a_fipspl** | Place FIPS (current) | Char | 5 |
| **a_geocode** | Tabulation state FIPS \|\| Tabulation county FIPS \|\| tabulation census tract | Char | 11 |
| **a_geoqual** | Quality of lat/long | Num | 8 |
| **a_latitude** | Block's internal point latitude, 6 implied decimals | Num | 8 |
| **a_longitude** | Block's internal point longitude, 6 implied decimals | Num | 8 |
| **a_ssccc** | Current state-county FIPS | Char | 5 |
| **a_st** | Current state FIPS | Char | 2 |
| **a_st_tab** | Tabulation state FIPS | Char | 2 |
| **a_taz** | Traffic Analysis Zone (current) | Char | 8 |
| **a_tract** | Current census tract | Char | 6 |
| **a_wib** | Workforce Investment Board area (current) | Char | 6 |
| **galid** | Unique GAL address ID: 'A' \|\| GAL vintage when added \|\| sequential integer | Char | 29 |
| **geovtg** | Geo-vintage for current geography | Num | 8 |

## 4.5 NOTES

# Chapter 5.
# Individual Characteristics File (ICF)

## 5.1 OVERVIEW

The National Individual Characteristics File (ICF) contains one record for every person who is ever employed in any LEHD state over the time period spanned by states's unemployment insurance records, conditional on participation in the LEHD program (the equivalent file for persons employed at some point by the federal government in positions covered by OPM is constructed independently, and documented separately). It consolidates information from multiple input sources on gender, age, place-of-birth, race, ethnicity, and education. Additional information on yearly place-of-residence since 1999 is also available. Information on age, gender, place-of-birth, race, ethnicity, and education is imputed ten times when missing.

### 5.1.1 Details of the Construction of the ICF Variables

#### 5.1.1.1 Overview

Each variable on the new ICF is derived from at least one of the following three files: the PCF, the HCEF, or the SCEF. The PCF is the Person Characteristics file, which is our conduit to Census' version of the SSA Numident file. The Numident contains information recorded from every transaction with SSA, including the initial application. The HCEF is the Hundred Percent Census Edited File and the SCEF is the Sample Census Edited File, commonly known as the short and long form respectively. Response variables on the HCEF and the SCEF generally begin with a Q and their corresponding flag begins with an F.

Many variables such as date of birth have multiple sources, but some such as education only have one source. If more than one source exists, one is designated as the primary and the other is the secondary. If the variable is missing then the standard SAS missing value for that type is used, a single blank space for character variables and a dot for numeric.

#### 5.1.1.2 Imputation process for demographic variables

The three source data files follow a monotone data pattern. This type of data pattern allows us to implement a hierarchical approach. We start with the variables with the least amount of missing data, the variables sourced from the PCF. In stage A we complete (replace a missing value with an implicate (the actual value returned from an imputation)) DOB, gender, and POB. With these variables completed we move on to the HCEF (stage B) and impute race and ethnicity conditional on the DOB, gender, and POB values from stage A. In the final stage we complete education conditional on the variables imputed in both stage A and B. By looking at the percent column in the table above you can get an idea of the amount of completed data for each variable. For example POB has about 5% completed data, while education has about 88% completed or 12% as reported (see Table 5.1.1.3 on page 5-2).

Table 5.1: Distribution of data sources for the ICF

| **PCF** | **HCEF** | **SCEF** | Percent |
|---|---|---|---|
| in | in | in | 12% |
| in | in | not | 61% |
| in | not | not | 22% |
| not | not | not | 5% |

#### 5.1.1.3 Updating records in the LEHD Production process

The full-information update is occassionally done by LEHD Research staff (at least once upon a state joining the LEHD Program). During regular production, workers not present on the (previous quarter's) National ICF show up every quarter (mostly new entrants to the US labor force in partner states). Thus, during the production process, a "new worker impute" is performed, using less information (and less computational resources) than the full-information impute. End-of-quarter processing unduplicates the state-level updates, and adds the new workers to the National ICF for the next production cycle.

Thus, the National ICF available to researchers in this snapshot will contain both records from one or more full-information process runs, and separate reduced-information runs at the state level. The SOURCE and VINTAGE variables identify the source of each record.

**Quantifying cumulative updates.** Across all available states, and cumulatively over a 4-quarter period, only about 0.03% of all PIKs are added by the state-level process. This fraction is somewhat higher among observations with imputed values (lag in availability of updated PCF, which itself has a lag with respect to the first-time appearance of workers on the labor market).

- Overall in the base data, about 10% of PIKs have a age or sex or POB impute. Among the updated (unique) records, this proportion is 35%.
- Overall in the base data, slightly more than 30% of PIKs have a race or ethnicity impute. Among the updated (unique) records, this proportion is slightly more than 55%.
- Overall in the base data, about 92% of PIKs have a education impute. Among the updated (unique) records, this proportion is over 99%.

**Overlap between state-level updates.** Because each state-level process identifies a PIK with missing data separately, workers who (over the course of the 4 quarters analyzed) appear in multiple states have independent imputes in each state. Naturally, these imputes are not identical. Over the four quarters, about 2% of PIKs appear in at least one other state (about 3.5% of those appear in more than 2 states). All imputes condition on the same type of information (all new workers, by definition, do not have a work history), and only one (randomly selected) record (and its imputes) is retained when updating the National ICF at the end of the production cycle.

#### 5.1.1.4 Place of residence imputation

Place of residence information on the ICF is derived from the StARS (Statistical Administrative Records System), as provided historically to LEHD as the Composite Person Record (CPR), as well as other newer data records.The vast majority of the individuals found in the UI wage records have information in these data on the place of residence down to the exact geographical coordinates. However, in slightly more than 1 percent of all cases the geography information is incomplete or missing. The QWI estimation relies on completed place of residence information. Because this information is a critical conditioning variable in the Unit-to-Worker Impute (U2W) (chapter 10) imputation model, all missing residential addresses are imputed. For a more detailed description of the place-of-residence imputation process, please see Vilhuber and McKinney (2014).

### 5.1.2 Variable Details

**Imputation flags** Every variable has a corresponding imputation flag variable that identifies the status (observed or imputed) of the main variable. The impute flag name is always of the format <varname>_imputed. For example for DOB the flag is named DOB_imputed.

- All impute flags can take on the following values:
  - "1" = "observed"
  - "2" = "imputed"
  - "3" = "imputed but not replaced, implicates 1-10 on implicate file"

  The third value for the imputation flag is assigned when the observed value of <varname> fails consistency checks, and is deemed implausible, for instance, when observed age at the beginning of the first quarter of labor market activity is less than 12 or greater than 85.

**Source flags** Every variable also has a corresponding flag variable that tells the user the source and status (reported or missing) of the main variable. The flag name is always of the format <varname>_flag. For example for DOB the flag is named DOB_flag. Two flag values are reserved: 0 indicates an "as reported" value and 9 indicates a missing value. These variables are primarily of use for internal processing, and only available in the RDC on the `icf_us_nonworker` (Section 5.3.6) file.

**Date of Birth**

- Primary Source File: PCF
  - Variables: DOBYYO, DOBMMO, and DOBDDO
- Secondary Source File:HCEF
  - Variables: QDB and FDB
- Output Variable Name: DOB
- Variable Construction: The year, month, and day variables from the PCF and HCEF are first cleaned in preparation for conversion to a valid SAS date. We begin by marking as ineligible for further processing any values that contain less than 4 numeric characters (0-9). Next, each variable still eligible is converted to numeric. The year value is checked first and if a valid year (year > current year - 126, which is a max age of 125) is present then processing continues. If month is between 1 and 12 and day is between 1 and 31 inclusive then a date is potentially complete to the day (our finest resolution). Each type is processed separately: if year only is present then the month and day are imputed, if year and month are present then the day is imputed and finally if all 3 are available then the day is checked to insure it is valid and if not it is replaced with the closest valid value (28, 29, or 30). The end result is a SAS date for each HCEF and PCF value with a valid year. In the final step, the information is combined with the PCF information taking precedence unless the PCF is not available or the HCEF is clearly superior (valid year, month, and day reported).
- Notes:
  - The SAS functions year(), month(), and day() can always be used to create a Gregorian calendar year of birth from the SAS date.
  - When calculating age, please use the following formula: age=(reference SAS date - DOB)/365.2425).

  - SAS does not consistently handle non-integer DOB values, which is a known issue, see http://support.sas.com/kb/24/808.html.

- DOB values:
  - Date of birth will be stored as a SAS date on the file. The SAS System stores date values as an offset in days from January 1, 1960. (SAS numeric 4)

- DOB_flag Values:
  - 0=PCF DOB valid and complete non-corrected
  - 1=PCF DOB valid and complete once day is corrected
  - 2=PCF DOB valid year and month, day is imputed
  - 3=PCF DOB valid year, month and day are imputed
  - 4=HCEF DOB of type 0 replaces a missing or type 2,3 PCF DOB
  - 5=HCEF DOB of type 1 replaces a missing or type 2,3 PCF DOB
  - 9=DOB missing

### Gender

- Primary Source File: PCF
  - Variable: gender
- Secondary Source File: HCEF
  - Variables: qsex and fsex
- Output Variable Name: gender (internal) sex (production)
- Variable Construction: PCF gender takes precedence over HCEF gender. HCEF qsex is used when PCF gender is missing and qsex is either reported or imputed based on the first name of the respondent.
- sex values:
  - M=Male
  - F=Female
- sex_flag values:
  - 0=PCF sex is a M or F
  - 1=HCEF sex (PCF sex is not a M or F and fsex is a 0 (reported) or a 1 (allocated based on first name))
  - 9=gender missing

### Place of Birth

- Primary Source File: PCF
  - Variables: POBST and POBFIN
- Secondary Source File: SCEF
  - Variables: qpobst and fpob
- Output Variable Name: POB
- Variable Construction: The PCF and HCEF variables are passed through formats, assigning each country code to either a new standardized country code or region. The individual country codes represent the top 23 immigrant source countries (including Puerto Rico) among all PIK records with a valid POBST and POBFIN and at least one quarter of positive earnings. Together, over 70% of the foreign born emigrated from one of the 23 source countries. In addition, the list contains every source country with at least 1% of the U.S. foreign born population.
- POB values:

  - A = US or territory (not Puerto Rico)
  - B = Mexico
  - C = Philippines
  - D = Vietnam
  - E = India
  - F = Germany
  - G = Puerto Rico
  - H = El Salvador
  - I = Cuba
  - J = United Kingdom
  - K = Canada
  - L = China
  - M = South Korea
  - N = Taiwan
  - O = Guatemala
  - P = Japan
  - Q = Haiti
  - R = USSR Core
  - S = Jamaica
  - T = Columbia
  - U = Poland
  - V = Iran
  - W = Dominican Republic
  - X = Italy
  - Y = Former Socialist Europe
  - Z = Western Europe
  - 1 = Central Asia
  - 2 = South East Asia
  - 3 = Middle East and North Africa
  - 4 = Caribbean
  - 5 = Central America
  - 6 = South America
  - 7 = Africa
  - 8 = Oceania

- POB_flag values:
  - 0=PCF POB is valid and complete
  - 1=HCEF POB is valid and complete as reported
  - 9=POB missing (Born abroad of an unknown country are included here, POBFIN=* and POBST=“ " are removed)

### Race

- Primary Source File: HCEF
  - Variables: imprace, frace, and fimprace
- Output Variable Name: race
- Variable Construction: A collapsed version of imprace is the primary source. This variable was chosen after an exhaustive analysis of both the recorded responses from the HCEF (inrace1-inrace21) and the variables qrace1-qrace8 that capture the 8 “best" responses. First, the inrace variables were shown to be mapped sensibly into the

qrace variables. A variable was created using qrace, but the values were consistent with imprace, once processing as described in the flags was applied. This rendered the variable constructed from qrace obsolete and imprace was used directly. The interaction of the flag variables frace and fimprace is represented in race_flag. Generally when the Census has information not available to our imputer, the HCEF edit or allocation was retained. For example, if a person's race response is missing but at least one other member of the household reports a valid race, then the allocation is retained. However, this rule was not blindly applied, the quality of the allocation was confirmed using bestrace from the Numident. The correspondence with bestrace must be relatively high for the allocation to be retained. No hot (cold) deck allocations were retained.

- race values:
  - 1=White Alone
  - 2=Black or African American Alone
  - 3=American Indian or Alaska Native Alone
  - 4=Asian Alone
  - 5=Native Hawaiian or Other Pacific Islander Alone
  - 7=Two or More Race Groups
- race_flag values:
  - 0=(frace=0 and fimprace=0) race is as reported
  - 1=(frace=0 and fimprace=1) For multiple race respondents, the write-in “some other" race value is dropped and the checkbox (valid) race value is retained
  - 2=(frace=0 and fimprace=4) For respondents with only a “some other" race value, a new valid race was allocated from within the household
  - 3=(frace=1 and fimprace=0) Code changed through a consistency check
  - 4=(frace=1 and fimprace=1) For multiple race respondents, the code changed through consistency check write-in “some other" race value is dropped and the checkbox (valid) race value is retained
  - 5=(frace=3 and fimprace=0) The classified from race response in the Hispanic question value is retained
  - 6=(frace=3 and fimprace=4) The classified from race response in the Hispanic question “some other" race value is allocated a new valid race from within the household.
  - 7=(frace=4 and fimprace=0) The allocated from within the household value is retained
  - 8=(frace=4 and fimprace=1) The allocated from within the household multiple race value is adjusted. The write-in “some other" race value is dropped and the checkbox (valid) race value is retained.
  - 9=race is missing
  - 10=(frace=4 and fimprace=4) The allocated “some other" race value is replaced with a non “some other" race value from another member of the household.

### Ethnicity

- Primary Source File: HCEF
  - Variable: qspan and fspan
- Output Variable Name: ethnicity
- Variable Construction: The variable qspan is passed through a format to assign the 3 digit codes to a simple Hispanic or non-Hispanic. See the variable flag for details of the values retained. The values retained were determined using a similar logic as the race variable. The main exception is that some hot deck values are retained. In this case, the surname hot deck is retained due to its relatively close correspondence with bestrace.
- ethnicity values:
- N=Not Hispanic or Latino
- H=Hispanic or Latino
- ethnicity_flag values:

  - 0=as reported
  - 1=Multiple responses were a given a unique Hispanic or non-Hispanic code
  - 2=Assigned Hispanic from the race code
  - 3=Allocated from within the household
  - 4=Allocated using a hot deck conditioning on surname
  - 9=ethnicity is missing

**Education (educ_c)**

- Primary Source File: SCEF
  - Source Variable: qhigh and fhigh
- Output Variable Name: educ_c
- Variable Construction: The SCEF qhigh values were collapsed using a format. If present, the variable DOB was used to calculate the respondents age on April 1, 2000. If the age was greater than or equal to 25 then the as reported education value was retained. EDUC_C is derived from EDUC_F, which is not available on Production or RDC files.
- educ_c values: (values from the full education coding are in parenthesis):
  - 1 = Less than high school (1-8)
  - 2 = High school or equivalent, no college (9)
  - 3 = Some college or Associate degree (10-12)
  - 4 = Bachelor's degree or advanced degree (13-16)
- educ_c_flag values:
  - 0=as reported education, DOB available, and calculated age greater than or equal to 25.
  - 9=education missing

### 5.1.3 Changes in this Snapshot

**Updated data** There has been no change to the imputation models that complete the data in the ICF. New workers having entered the LEHD covered workforce in the time period covered have been added, and where necessary, lookups and imputations for complete demographic details performed as outlined in the main document.

## 5.2 DATA CITATION

U.S. Census Bureau. 2016. *Individual Characteristics Files (ICF) in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

## 5.3 DATA SET DESCRIPTIONS

### 5.3.1 Unique record identifier

The unique record identifier within each ICF file is the **P!** (**P!**)IK.

### 5.3.2 Naming scheme

There are 10 files in the ICF/ICFT26 group: SAS datasets with zero observations are attached to this document:[1]

- icf/us/icf_us_implicates_age_sex_pob.sas7bdat
- icf/us/icf_us_implicates_education.sas7bdat
- icf/us/icf_us_implicates_race_ethnicity.sas7bdat
- icf/us/icf_us_nonworkers.sas7bdat
- icf/us/icf_us.sas7bdat
- icft26/us/icf_us_addresses.sas7bdat

### 5.3.3 Data location

The files are stored in two main directories:

```
icf/us/       for most files
icft26/us     for files with Title 26 protected content
```

[1] Also visible on the attachment tab - Adobe Reader may be required.

### 5.3.4 Main dataset: ICF_us

This is the core dataset, containing all observed non-FTI and the first implicate for imputed variables.

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 5.2: ICF_US: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **DOB** | Date of birth | Num | 4 |
| **DOB_imputed** | Imputation status for DOB | Char | 1 |
| **PIK** | Protected Identification Key | Char | 9 |
| **POB** | Place of birth | Char | 1 |
| **POB_imputed** | Imputation status for POB | Char | 1 |
| **educ_c** | Highest educational attainment (age 25+) | Char | 1 |
| **educ_c_imputed** | Imputation status of educ_c | Char | 1 |
| **ethnicity** | Ethnicity | Char | 1 |
| **ethnicity_imputed** | Imputation status of ethnicity | Char | 1 |
| **race** | Race | Char | 1 |
| **race_imputed** | Imputation status for race | Char | 1 |
| **sex** | Gender | Char | 1 |
| **sex_imputed** | Imputation status for sex | Char | 1 |
| **source_process** | us=created by NICF process, if state abbreviation=new worker process in state | Char | 2 |
| **vintage** | Date and Time of File Creation | Char | 13 |

### 5.3.5 Utility dataset (view): ICF_us_wide and NICF_us_wide

This is a SAS view (views do not work in Stata). For the utility of users wishing a wide file, this view combines all variables on all implicates (<varname>[n]) and the variables from the core ICF file (<varname>) into single dataset. Note that a view performs the merge "on the fly". The only difference between the two views is the naming of the sex/gender variables ("sex" on `icf_us_wide`, "gender" on `nicf_us_wide`). Only the variables from `icf_us_wide` are listed below. In general, researchers should use `icf_us_wide.sas7bvew`. If needed, users can customize the view by extracting and modifying the program `create_icf_wide.sas` (see Section 5.4.3).

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 5.3: ICF_US_WIDE: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **DOB** | Date of birth | Num | 5 |
| **DOB1** | date of birth (sas date) | Num | 5 |
| **DOB2** | date of birth (sas date) | Num | 5 |
| **...** | | | |
| **DOB10** | date of birth (sas date) | Num | 5 |
| **DOB_imputed** | Imputation status for DOB | Char | 1 |
| **PIK** | Protected Identification Key | Char | 9 |
| **POB** | Place of birth | Char | 1 |
| **POB1** | Place of birth Implicate 1 | Char | 1 |
| **...** | | | |
| **POB10** | Place of birth Implicate 10 | Char | 1 |
| **POB_imputed** | Imputation status for POB | Char | 1 |
| **_merge** | base \|\| age \|\| race \|\| educ | Char | 4 |
| **educ_c** | Highest educational attainment (age 25+) | Char | 1 |
| **educ_c1** | Highest educational attainment (age 25+) Implicate 1 | Char | 1 |
| **...** | | | |
| **educ_c10** | Highest educational attainment (age 25+) Implicate 10 | Char | 1 |
| **educ_c_imputed** | Imputation status of educ_c | Char | 1 |
| **ethnicity** | Ethnicity | Char | 1 |
| **ethnicity1** | Ethnicity Implicate 1 | Char | 1 |
| **...** | | | |
| **ethnicity10** | Ethnicity Implicate 10 | Char | 1 |
| **ethnicity_imputed** | Imputation status of ethnicity | Char | 1 |

(cont.)

**Table 5.3 (cont.): ICF_US_WIDE: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **race** | Race | Char | 1 |
| **race1** | Race Implicate 1 | Char | 1 |
| **...** | | | |
| **race10** | Race Implicate 10 | Char | 1 |
| **race_imputed** | Imputation status for race | Char | 1 |
| **sex** | Sex | Char | 1 |
| **sex1** | Sex Implicate 1 | Char | 1 |
| **...** | | | |
| **sex10** | Sex Implicate 10 | Char | 1 |
| **sex_imputed** | Imputation status for sex | Char | 1 |
| **source_process** | us=created by NICF process, if state abbreviation=new worker process in state | Char | 2 |
| **vintage** | Date and Time of File Creation | Char | 13 |

**Usage:** A SAS view is read the same way a regular SAS dataset is read:

```
libname icf "&snapshot./s2014/icf/us" access=readonly;
data mysample;
set icf.icf_us_wide (where=(substr(pik,1,2)='01'));
/* further processing steps */
run;
```

### 5.3.6 Auxiliary dataset: ICF_us_nonworkers

This dataset contains observed values from all basedata, for records that were not completed in the first vintage of the National ICF research file. Due to updates from both the full-information and new-worker processes, there is overlap between this file and the universe of icf_us. This file contains no imputes, and only those variables from the basedata necessary for processing the full-information process.

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 5.4: ICF_US_NONWORKERS: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **DOB** | date of birth (sas date) | Num | 5 |
| **DOB_flag** | date of birth source and quality | Num | 4 |
| **PIK** | PIK - Protected Identification Key | Char | 9 |
| **POB** | country/region of birth | Char | 1 |
| **POB_flag** | POB source and quality | Num | 4 |
| **educ_c** | self reported educ (age>=25 on 4/1/2000) | Char | 1 |
| **educ_c_flag** | education source and quality | Num | 4 |
| **ethnicity** | self reported Hispanic (H/N) | Char | 1 |
| **ethnicity_flag** | ethnicity source and quality | Num | 4 |
| **gender** | male (M) or female (F) | Char | 1 |
| **gender_flag** | gender source and quality | Num | 4 |
| **pcf_race** | recoded Numident bestrace | Char | 1 |
| **race** | self reported race | Char | 1 |
| **race_flag** | race source and quality | Num | 4 |
| **rowid** | Source data available | Char | 1 |
| **stage** | First vars impute needed | Char | 1 |
| **type** | Source PCF HCEF SCEF | Char | 3 |

### 5.3.7 Age, sex, and place-of-birth implicates: ICF_us_implicates_age_sex

The first implicates for date of birth, sex, and place-of-birth are stored on the main ICF file as DOB, SEX, and POB. Imputed values are flagged by the appropriate flag. Other implicates are found in this file, and can be merged on when required.

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 5.5: ICF_US_IMPLICATES_AGE_SEX_POB: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **DOB1** | date of birth (sas date) | Num | 5 |
| **DOB2** | date of birth (sas date) | Num | 5 |
| **...** | | | |
| **DOB10** | date of birth (sas date) | Num | 5 |
| **PIK** | Protected Identification Key | Char | 9 |
| **POB1** | Place of birth Implicate 1 | Char | 1 |
| **...** | | | |
| **POB10** | Place of birth Implicate 10 | Char | 1 |
| **sex1** | Gender Implicate 1 | Char | 1 |
| **...** | | | |
| **sex10** | Gender Implicate 10 | Char | 1 |
| **source_process** | us=created by NICF process, if state abbreviation=new worker process in state | Char | 2 |
| **vintage** | Creation date of record | Char | 13 |

### 5.3.8 Education implicates: ICF_us_implicates_education

The first implicate is stored on the main ICF file as EDUC_C. Imputed values are flagged by the appropriate flag. Other implicates are found in this file, and can be merged on when required.

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 5.6: ICF_US_IMPLICATES_EDUCATION: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **PIK** | Protected Identification Key | Char | 9 |
| **educ_c1** | Highest educational attainment (age 25+) Implicate 1 | Char | 1 |
| **...** | | | |
| **educ_c10** | Highest educational attainment (age 25+) Implicate 10 | Char | 1 |
| **source_process** | Source of Impute | Char | 2 |
| **vintage** | Vintage of Impute | Char | 13 |

### 5.3.9 Race and ethnicity implicates: ICF_us_implicates_race_ethnicity

The first implicates are stored on the main ICF file as RACE and ETHNICITY. Imputed values are flagged by the appropriate flag. Other implicates are found in this file, and can be merged on when required.

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 5.7: ICF_US_IMPLICATES_RACE_ETHNICITY: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **PIK** | Protected Identification Key | Char | 9 |
| **ethnicity1** | Ethnicity Implicate 1 | Char | 1 |
| **...** | | | |
| **ethnicity10** | Ethnicity Implicate 10 | Char | 1 |
| **race1** | Race Implicate 1 | Char | 1 |
| **...** | | | |
| **race10** | Race Implicate 10 | Char | 1 |
| **source_process** | Source of Impute | Char | 2 |
| **vintage** | Vintage of Impute | Char | 13 |

### 5.3.10 Title 26 information: ICF_us_addresses

FTI has been removed from the core ICF, and stored separately. Note that in the RDC network, this file is stored under a separate set of permissions, and if users require access to this information, need to request access to an additional group. In contrast to previous snapshots, only a single impute is provided.

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 5.8: ICF_US_ADDRESSES: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **address_year** | Year of address record - worker worked in this year | Num | 3 |
| **county_live** | FIPS State(2) \|\| FIPS County (3) as of address_year | Char | 5 |
| **countyliveimputed** | County of Residence imputation flag | Char | 1 |
| **flag_distance** | Years away from observed CPR value (edit flag) | Num | 3 |
| **flag_latlong** | Flag quality of latitude/longitude of residence | Num | 3 |
| **huid** | HUID - Admin Record HUID | Char | 35 |
| **latitude_live** | Latitude of residence, 6 implied decimal places | Num | 8 |
| **longitude_live** | Longitude of residence, 6 implied decimal places | Num | 8 |
| **pik** | PIK - Protected Identification Key | Char | 9 |
| **source** | Source process (state name=ICF for that state) | Char | 3 |
| **vintage** | Vintage in which record was created | Char | 13 |

## 5.4 HELPFUL PROGRAMS

The following programs might be found to be useful when using the data.

### 5.4.1 Recombining T26 data with the core ICF

The following program allows users to combine the Title 26 variables with the core ICF. This program was used in slightly modified form for quality assurance during the preparation of the data for the RDC environment.

```
/* Time-stamp: <07/05/03 23:49:08 vilhuber> */
/* $Id: 02.02.combine_icf_t26.sas 121 2007-05-04 12:18:17Z vilhu001 $ */



%macro combine_icf_t26(state=,inlib=WORK,int26=WORK);

libname INLIB "/mixedtmp/lehd/s2004/icf/&state./";
libname INT26 "/mixedtmp/lehd/s2004/icft26/&state./";
libname INPUTS (&inlib., &int26.);

libname ORIG "/mixedtmp/lehd2/s2004_obsolete/icf_commingled/&state./" access=readonly;

proc sort data= ORIG.icf_&state out= icf_orig(compress=yes);
by pik;
run;

data work.merged(sortedby=pik state);
merge INPUTS.icf_&state._t26  INPUTS.icf_&state.;
by pik state;
run;

proc contents data=icf_orig;
run;
proc contents data=work.merged;
run;

*proc compare data=icf_orig briefsummary compare=work.merged;
*run;
%mend;


/* example - this works for all states */
libname temp '/temporary/saswork1/snapshot';
options mprint symbolgen;
%combine_icf_t26(state=al,inlib=INLIB,int26=INT26);
```

### 5.4.2 Selecting a random subsample of persons

The following program allows users to select a random sample of approximately one percent of individuals on the ICF. It relies on the fact that the first two characters of the PIK are approximately uniformly distributed on $[00, 99]$. Note that 'AA' is a valid value for the first two characters and denotes individuals for whom no valid SSN was on file. Occurrence of such "pseudo-PIKs" varies by state.

```
%let syear=2014;
libname INLIB "&snapshot./s&syear./icf/us/";

data my_icf;
   set INLIB.icf_us(where=(substr(PIK,1,2)='01'));
run;
```

Alternatively, you can use the method described in Section 3.5, which does not require access to the ICF.

### 5.4.3 Creating a ICF wide file

This view combines all variables on all implicates (<varname>[n]) and the variables from the core ICF file (<varname>) into single dataset. Note that a view performs the merge "on the fly". The only difference between the two views is the naming of the sex/gender variables ("sex" on `icf_us_wide`, "gender" on `nicf_us_wide`). The contents of `icf_us_wide` are described in Section 5.3.5.

```
/* Create two wide files from the production NICF snapshot files */

options msglevel=i ls=150 ps=10000 mprint mlogic symbolgen obs=max;


libname INPUTS ".";

/* macro is defined here

   Sample call:
   %create_icf_wide(icf_dir=/path/to/icf,research=yes);

*/

%macro create_icf_wide(icf_dir=,research=yes);

%let prefix=icf;
%if ( "&research" = "yes" ) %then %let prefix=nicf;

/* The RESEARCH version uses gender, the PRODUCTION version uses SEX. Don't ask. */
data INPUTS.&prefix._us_wide(keep=PIK DOB DOB_imputed
                POB POB_imputed race race_imputed ethnicity
                ethnicity_imputed educ_c educ_c_imputed
                DOB1-DOB10
%if ( "research" = "yes" ) %then %do;
gender gender_imputed
gender1-gender10
%end; %else %do;
sex sex_imputed
sex1-sex10
%end;
POB1-POB10 race1-race10 ethnicity1-ethnicity10 educ_c1-educ_c10
                vintage source_process _merge

%if ( "&research" = "yes" ) %then %do;
                      rename=(
              sex  = gender
      sex_imputed=gender_imputed
                              sex1 = gender1
                              sex2 = gender2
                              sex3 = gender3
                              sex4 = gender4
                              sex5 = gender5
                              sex6 = gender6
                              sex7 = gender7
                              sex8 = gender8
                              sex9 = gender9
                              sex10= gender10)
%end; /*end research condition*/

sortedby=pik)

 /view=INPUTS.&prefix._us_wide;

  merge "&icf_dir./icf_us"(in=a)
        "&icf_dir./icf_us_implicates_age_sex_pob"(in=b drop=vintage source_process)
        "&icf_dir./icf_us_implicates_race_ethnicity"(in=c drop=vintage source_process)
        "&icf_dir./icf_us_implicates_education"(in=d drop=vintage source_process);
     by pik;

  array DOB_impl DOB1-DOB10;
  array sex_impl sex1-sex10;
  array POB_impl POB1-POB10;
  array race_impl race1-race10;
  array ethnicity_impl ethnicity1-ethnicity10;
  array educ_c_impl educ_c1-educ_c10;

  length _merge $4;
  %if ( "&research" = "yes" ) %then %do;
  label sex_imputed = "Imputation status for gender";
%end; /*end research condition*/
  %else %do;
          label sex  = "Sex"
        sex1 = "Sex Implicate 1"
        sex2 = "Sex Implicate 2"
        sex3 = "Sex Implicate 3"
        sex4 = "Sex Implicate 4"
        sex5 = "Sex Implicate 5"
        sex6 = "Sex Implicate 6"
        sex7 = "Sex Implicate 7"
        sex8 = "Sex Implicate 8"
        sex9 = "Sex Implicate 9"
        sex10= "Sex Implicate 10"
;
  %end;
     label _merge="base || age || race || educ";

  _merge = put(a,1.) || put(b,1.) || put(c,1.) || put(d,1.);

  if a=1 then do;

     if DOB_imputed="1" then do over DOB_impl;
       DOB_impl=DOB;
     end;

     if sex_imputed="1" then do over sex_impl;
       sex_impl=sex;
     end;

     if POB_imputed="1" then do over POB_impl;
       POB_impl=POB;
     end;

     if race_imputed="1" then do over race_impl;
       race_impl=race;
     end;

     if ethnicity_imputed="1" then do over ethnicity_impl;
       ethnicity_impl=ethnicity;
     end;

     if educ_c_imputed="1" then do over educ_c_impl;
       educ_c_impl=educ_c;
     end;


     output INPUTS.&prefix._us_wide;
  end;
run;

data view=INPUTS.&prefix._us_wide;
   describe;
run;

proc contents data=INPUTS.&prefix._us_wide;
proc print data=INPUTS.&prefix._us_wide(obs=100);
run;
%mend;
```

## 5.5 NOTES

# Chapter 6.
# Office of Personnel Management files (OPM)

## 6.1 OVERVIEW

The present chapter describes how federal workers are added to the QWI infrastructure. The core data are provided by OPM. We highlight the differences between the structure and content of the data provided by OPM and the data provided by state UI systems, and the efforts undertaken to make the data comparable.

The OPM data create some challenges. In contrast to the data from the state UI systems, which record cumulative employment over a quarter, OPM data are provided as a database extract, with a true point-in-time stock of employees at the end of a calendar year quarter, and a separate file providing for information on status changes. Whereas the UI systems record cumulative earnings, the OPM system only records the nominal annual salary, plus an indicator of whether or not a particular employee is full-time, part-time, or seasonal; neither system records actual hours worked. Finally, work location is not collected in the same manner as in the QCEW, and industry is not collected at all.

We have implemented solutions for all of these shortcomings. Federal workplaces do report their workplace employment in the QCEW. We leverage this information both to address the absence of precise workplace location in the OPM-provided data, and to assess coverage.

### 6.1.1 Requesting Access to OPM Data in the FSRDC

Under an MOU between the Office of Personnel Management (OPM) and the Census Bureau, the Census Bureau will incorporate information on the federal workforce into the Census Bureau's data infrastructure, in order to "*improve economic and demographic censuses, surveys, and intercensal population estimates.*" Research using these files is intended to further support the "*Master Address File Program, current demographic and economic survey and census operations.*" The microdata are protected unter Title 13, U.S.C.

> All research proposals that request access to the OPM data in the FSRDC must do so for the same purpose as just outlined. For instance, it is not feasible to request access to OPM data for the sole purpose of studying the federal workforce.

We note that OPM also releases quarterly data on the federal government's workforce at `http://www.fedscope.opm.gov`, and allows access to individual-level (de-identified) data underlying the FedScope data at `http://www.opm.gov/data/Index.aspx?tag=FedScope`. The data provided to the Census Bureau are extracted from the same Enterprise Human Resources Integration-Statistical Data Mart (EHRI-SDM) that feeds FedScope and the raw data at the above location. Proposals requesting access to the OPM data in the FSRDC system should address both the need for data not otherwise available, and the specific benefit to the Census Bureau, as noted above.

## 6.1.2 Data Sources and Definitions

### 6.1.2.1 Office of Personnel Management input data files

The OPM data provided to LEHD is composed of four types of files:

- Dynamics file: A personnel action file describing personnel actions for federal workers that took place during the quarter (and sometimes, took place in previous quarters but didn't show up in the file until later). In addition to basic characteristics of the workers largely included also on the status file (described below), the dynamics file records personnel actions for each federal worker. Personnel actions include accessions, separations, promotions, movements between different work schedules, adjustments in locality or basic pay, etc. The date of each action is recorded at daily precision.
- Status file: A status 'snapshot' of the federal workforce on a particular date (the last day of the last pay period in the calendar year quarter). Most of the variables on the status and dynamics file overlap, but not all. A worker will appear in the status file but not in the dynamics file if no personnel actions took place for this worker in that quarter. A worker will appear in the dynamics file, but not the status file, if that worker's attachment to the federal workforce was terminated during the quarter. Other, more complex situations, may also occur. Lags sometimes occur between actions described in the Dynamics file and their reflection in subsequent Status files. This will be addressed dynamically during processing.
- Standard Code Table (SCT) file: The Standard Code Table file is a lookup table for values in the fields in the Dynamics and Status files.
- Point of information (POI) file: A personnel office address file that gives a street address for the personnel officer contact.

In addition, a Duty Station File is available from the OPM website, mapping duty station codes to CBSA. This is used for QA purposes. Detailed information on data elements can be found at http://www.opm.gov/feddata/guidance.asp. Note that not all elements are available on the files provided to LEHD. Overall, OPM provides data on 543 agencies (402 Cabinet Level Agencies from 18 departments, 141 independent agencies).[1]

The raw input files provided by OPM to LEHD are not made available to outside researchers.

### 6.1.2.2 Available data and definitions

For the LEHD infrastructure, and to be compatible with the UI wage records, the key variable is quarterly earnings. We use the variable `totpay` (Total pay). All employment statistics are constructed for periods where positive earnings were received. `totpay` is computed by OPM, and includes

- basic pay
- locality adjustment
- supervisory differential
- retention allowance
- cost of living allowance (COLA)

We further use demographic data contained within the file, for tabulation of OPM-specific data, as well as to enhance other LEHD tabulations. Treatment of demographic data are described in Section 6.1.3.8.

### 6.1.2.3 Missing elements

For the LEHD production system, the critical elements missing are the realized quarterly earnings of employees, the roof-top address of the worker's workplace, and the industry coding of the agency. We address these issues in Section 6.1.3. Furthermore, while it is possible to identify part-time or intermittent workers from data fields, no hours

[1] Data derived from public-use OPM files, current as of Sept 2011.

are reported, and the totpay variable contains the full-year-equivalent earnings. We describe how this is addressed in Section 6.1.3.

#### 6.1.2.4 Problematic issues

**Personnel actions with effective dates in quarter $n$ may not appear in the Dynamics file until quarter $m > n$** Since agencies can delay Dynamics file submissions, and resubmit incorrect submissions up to two years later, a retrospective processing may be needed. We address this through a search of the entire Dynamics file history, and allow for revisions up to four quarters before the most current quarter being processed.

**Personnel actions that lead to termination of employment with a particular agency in quarter $n$ lead to no records (and thus no earnings) appearing in the Status file in quarter $n$** . The Status file only reflects point-in-time employment at the end of the quarter. Any termination of employment within the quarter means the worker has no record at that job in that quarter. We address this by identifying such situations from the Dynamics file, and adding employment records in quarter $n$. *(Not implemented in 2012Q2 release)*

**Award amounts not included in total pay** We use only the totpay variable to define (regular) earnings. To completely compute earnings, any award amount (awardamt) from the Dynamics file (e.g., cash bonus, separation incentive, student loan repayment, etc.) needs to be added. The newer OPM system (Enterprise Human Resources Integration (EHRI)) has a variable that computes such complete earnings automatically, but only since 2009. Furthermore, the LEHD MOU pre-dating the introduction of EHRI, that variable is not transmitted to LEHD. In 2006, approximately 30% of workers received such awards.

**Departmental reorganizations** In 2002, the Department of Homeland Security (DHS) was created as a Cabinet-level department by Congress, with an effective date of March 1, 2003.[2] Multiple transfers of agencies, including complex re-assignment of personnel and responsabilities, were implemented (Table 6.11), which complicate flows. QWI tabulations uses a flow-based approach to capturing such reorganizations in the SPF (chapter 9) as they relate to separations and hires, but by design will not capture partial splits and mergers (Benedetto et al. 2007). However, the large movement of agencies under the umbrella of the DHS seems to be captured.

#### 6.1.2.5 Exclusions

Certain agencies provide no data to OPM (U.S. Office of Personal Management 2012) and are thus excluded from the data universe for LEHD processing (see Table 6.7 on page 6-27). We also chose to exclude several agencies at this time due to particular processing restrictions:

- The concordance of the geographic location (county) of many Department of Defense bases was weak between OPM records and QCEW records. Until this matter can be resolved, all **Department of Defense (civilian)** employees are excluded from the QWI-OPM universe. (Department codes DD, AR, AF, NV)
- OPM, in its official publications, does not disclose the location (at the state level) of employees of **several security-related agencies**, other than that they may be working in the general Washington D.C. area. Because LEHD tabulations rely on sub-state geography, we have excluded these agencies from the QWI-OPM universe (see Table 6.8).
- The Federal Bureau of Investigation (FBI) provided Status file data to OPM until FY 2007, but did not provide Dynamics file data. Because we rely on the Dynamics file for critical timing information, all **FBI**-provided information is excluded from the universe for all periods. (Agency code: DJ02)

[2] Homeland Security Act, November 2002, see http://www.dhs.gov/xabout/history/editorial_0133.shtm

- The **State Department** no longer provides Status and Dynamics file data on workers in the **Foreign Service** after 2006Q2, and those workers are thus out of scope from that point onwards (drop in employment). Other State Department employees, however, are included.

## 6.1.3 Integration Methodology

### 6.1.3.1 Creating wage-record like files from OPM

In this section, we describe how the wage record files are created that match the structure of the UI files used in the rest of the LEHD system of files. In particular, we address creation of quarterly status snapshots (Section 6.1.3.2), adjustments of earnings (Section 6.1.3.4), and attaching detailed geographic and industry, consistent with QCEW coding, to duty stations (Section 6.1.3.5).

### 6.1.3.2 Selecting records

We first exclude data that are out-of-universe, suffer from known data quality issues, or other issues, as noted earlier (see Tables 6.7 and 6.8 as well as release notes). Once exclusions have been processed, a first pass through the data creates a "pseudo-UI" wage record: a record for any job (employment relation with an agency) during a quarter.

### 6.1.3.3 Adjusting records

We parse the dynamics files for each quarter, and for each employee, an additional determination is made as to whether the employee was employed by some agency during the quarter. Because the status file records only active jobs in the last pay period of the quarter, jobs that end within a quarter, but were active at the start of the quarter, are not reflected in the status file - in contrast to UI wage records, where separations are reflected as wage records in the quarter they occur. Some additional adjustments also occur.

**Impute missing dynamics records to improve consistency of dynamics and status file.** If the status file reports an individual is on unpaid status in-between consecutive accessions or recall actions reported on the dynamics file, a separation is imputed to have occurred either in the middle of the quarter, or halfway in-between the last accession and the end of the quarter, whichever is later. Conversely, if the status file reports an individual is on paid status in-between consecutive separation actions reported on the dynamics file, an appointment is generated in the middle of the quarter, or halfway in-between the last separation and the end of the quarter, whichever is later.

**Adjusting earnings.** Earnings are computed from `totpay`, which comprise basic pay, locality adjustments, supervisory differential, retention allowance, as well as COLA. `totpay` reflects a full-time, full-year equivalent salary, not actual earnings. When accessions or separations occur within a quarter, the exact number of days in the quarter that a worker was in pay status is computed, and stored as a fraction. Earnings are adjusted at a later stage (after processing of seasonal, intermittent, and part-time workers, see next section).

### 6.1.3.4 Processing seasonal, intermittent, and part-time workers

OPM data identify whether a worker is part-time (PT), seasonally, or intermittently (SI) employed. However, the data do not identify the actual number of hours during any time period, and do not directly identify the number of days worked during a quarter. About 6.2% of workers are classified as either part-time, seasonally, or intermittently employed.

Because the standard UI data records exclusively record the actual earnings (or wage-like payments) received by a worker, an adjustment is required.[3] In order to make OPM data compatible with UI earnings concept, we adjust the earnings for non-fulltime workers. For seasonally or intermittently employed workers, we compute the exact dates when a worker is in pay status throughout a calendar-year quarter from the dynamics file, and adjust the full-time, full-year equivalent earnings accordingly. For instance, if a worker was employed on a seasonal basis, and worked until April 15, then entered a non-pay status, returning to pay-status again on June 15, then the adjustment ratio for that worker is 30/91, and 33% of the full-time, full-year equivalent pay for that quarter is recorded as earnings on the pseudo-UI record.[4]

For workers identified as part-time workers, we impute hours worked during a quarter. Using a model based on Current Population Survey (CPS) and Decennial Census (DC) data, we condition on demographic characteristics, job characteristics, (employment history information: number of employers) and draw 10 imputes of hours worked. Actual quarterly earnings are then computed by multiplying the full-time, full-year pay rate by the ratio of imputed hours to potential quarterly hours.

#### 6.1.3.5 Attaching geographic and industry classification to OPM records

OPM records do not provide industrial classification (North American Industry Coding System (NAICS)) of the agency's activity. Furthermore, geographic precision of the agency's location is limited to the city the workerÂ´s dutystation is located in. While in principle, the QWI tabulations only require county-level precision, OnTheMap requires roof-top precision. In order to compensate for these two data shortcomings, the OPM agencies are matched to their corresponding QCEW reports. However, the OPM and QCEW share no common identification variables. We probabilistically match the two universes by name, higher-level geography, and other attributes to obtain a correspondence between the two sets of identifiers (OPM agency ID, and SEIN on the QCEW records. The following section describes how that matching is done. In an ideal world, with consistent naming and high-level geography (county) on both OPM and QCEW records, this would be a straightforward exercise with little if any uncertainty. Unfortunately, real world data is not ideal. Across the 50 states and the District of Columbia, the 543 agencies known to OPM expand to 647 name variations on QCEW. Some agencies are not reported in the QCEW with positive employment in the same county that OPM personnel records show active employees' duty stations. Others, in particular military bases, are recorded in different counties altogether. We describe our attempts at addressing them below.

#### 6.1.3.6 Selecting agency/establishment records

As mentioned, some agencies are excluded from the LEHD-OPM universe (see Section 6.1.3.2). These agencies must also be excluded as possible match candidates from the QCEW (see Table 6.8). We establish a master exclusion list, based on standardized OPM names. Agencies' records in the QCEW do not necessarily have time consistency or spatial consistency. Put differently, there are substantial variations in names of agencies and departments within any given state's historical QCEW records, and significant variations across states for the same agency. Thus, name matches are not exact, and probabilistic matching is used. [5] Records where the reliability of the probabilistic match is too low are clerically edited. All matches are databased, and are re-used for subsequent quarterly processing.

We note that in regular processing, the matches obtained (and validated) in previous periods are re-used. A database is maintained of all historical matches (SEIN and SEINUNIT assigned to specific agencies in each county), and a lookup against this database is performed prior to all name-matching. If the lookup is successful, and the establishment still exists in the corresponding QCEW data for that quarter, then no further matching is attempted.[6]

[3] Note, however, that UI wage records typically contain no information on hours, weeks, or days worked, and no information on part-time or intermittent status.

[4] Note that we assume that seasonal or intermittent workers are full-time when they do work.

[5] The Census Bureau uses SAS Data Quality Server for this purpose.

[6] Under certain conditions, the establishment need not exist in current QCEW records for the lookup result to be deemed valid.

An additional complication is that the level of detail (multi-unit breakouts) may differ between QCEW and OPM: QCEW may show more or fewer establishments in a particular county than OPM, for different states and different time periods.

The end result of this step are two lists: OPM agencies that are in-scope and define the universe that will be used as a baseline in all subsequent processing and reporting, and QCEW establishment records that are approximately equivalent in terms of the covered entities. The next step then consists in finding the QCEW reported establishment that corresponds to each reported agency, thus acquiring the QCEW record's industry classification and roof-top geocodes.

#### 6.1.3.7 Name-based matching to obtain SEIN and SEINUNIT

Because there are no common identifiers in the QCEW and OPM firm-level datasets, we resort to a probabilistic matching strategy that relies on exact and fuzzy matches by name, size, and location. To this end, we aggregate the compatible pseudo-UI records created in a previous step up to the department-agency-county level. These records, which are equivalent to the establishment-level records in the QCEW, are then matched to in-scope QCEW records, using a variety of matching criteria. The exact matching strategy is outlined in Table 6.10. Detailed results are available in a separate document.

The Duty Station code contains the city and county. We retain the county as a matching and blocking criterion. In general, we attempt to find the QCEW agency within the same county, but in some instances, there are persistent mismatches between OPM and QCEW counts within county, and the county criterion is relaxed, allowing agencies to match in other counties.

If the algorithm fails to find a likely establishment, a firm-level identifier (SEIN) is attached based on an impute that takes into account the employment-weighted distribution of establishments within the state (match passes 81-96, 4.8 percent). Note that for these records, the standard Unit-to-Worker Impute (U2W) algorithm in the LEHD infrastructure (Abowd et al. 2009) will be used to perform a probabilistic allocation of workplaces, and thus industry and geography, to jobs.

Finally, about 4.2 percent of matches are made during clerical review, either by visual inspection of candidate records, or through a series of custom algorithmic edits.

#### 6.1.3.8 Demographic information

OPM records contain demographic information for OPM workers. The information is provided each quarter, and the underlying personnel records are updated based on certain triggers. For the purposes of LEHD processing, the following data elements are required:

- Gender
- Age
- Race
- Ethnicity
- Education

However, some records are either incomplete (have missing information) or are not coded to the Office of Management and Budget (OMB) standard that LEHD uses currently (in particular, race and ethnicity from federal employees who were hired in their most recent position before 2006). We complete the information using two types of imputation models:

1. for variables with very few missing records, we derive the likelihood for the imputation model from the observed values for all ever-observed OPM workers, and construct a non-Bayesian draw from the likelihood;
2. for variables with high-levels of missingness, essentially race and ethnicity, we use a likelihood derived from the private-sector population data underlying the ICF, derived from 2000 Decennial information. The imputation conditions on reported non-OMB-compliant race and ethnicity, among other items, and jointly imputes OMB-compliant race and ethnicity combinations, and thus is better described as a probabilistic recode than a pure

impute. For workers with a recent personnel action that involved re-declaring race and ethnicity, no impute is necessary.

Impute rates are generally low, with the exception of OMB-compliant race and ethnicity:

| *Variable* | *Impute rate* |
|---|---:|
| Education (collapsed) | $< 5\%$ |
| Gender | $< 0.01\%$ |
| Date of birth | $< 0.01\%$ |
| Race and ethnicity | $< 61\%$ |

### 6.1.4 OPM-related files available to researchers

Once all data are integrated, the LEHD system creates files that correspond closely to the general LEHD Infrastructure files (see Section 6.3). Some caveats apply, however:

1. The processing programs used for OPM data may be of an older vintage than the most recent LEHD production processing used for the rest of the LEHD Infrastructure, due to the ongoing "preliminary" nature of the integration. In particular, in the S2014Snapshot, the OPM equivalent of the ECF and the QWI-SEINUNIT may differ from the structure as described in chapter 2) and chapter 7, respectively. The OPM-related files are provided as-is. It may be useful to consult older snapshot documents, if necessary.
2. The core "firm" identifier is an SEIN, not the agency code provided on the original OPM data files. In particular, this will lead to OPM employees of the same agency to appear to be employed by different entities in different states. This may not correspond to a researcher's intended definition of a government agency, but it does correspond closely to the definition of the employing entity in the UI wage records.

### 6.1.5 Changes in this Snapshot

OPM data on Federal workers were first added to the Snapshot with S2011. In the S2014 version of the snapshot, OPM data have been collected under a single directory. RDC users should be able to access these files by requesting a "OPM" dataset. Access to the OPM data do not require state permissions.

## 6.2 DATA CITATION

U.S. Census Bureau. 2016. *Office of Personal Management (OPM) files in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

## 6.3 DATA SET DESCRIPTIONS

### 6.3.1 Naming scheme

The OPM naming scheme is somewhat inconsistent with the remaining infrastructure files, reflecting the early-access nature of the data files. All files start with `opm`, followed by `_us` as the geographic indicator, and then the generic name of the file that is being supplied. This differs slightly from the naming conventions of UI-derived Infrastructure files. For instance, the EHF for OPM records is called:

```
opm_us_ehf.sas7bdat
```

and not `ehf_opm.sas7bdat`, and the ECF SEINUNIT file is called

```
opm_us_ecf_seinunit.sas7bdat
```

and not `ecf_opm_seinunit.sas7bdat`. Future snapshots should have a consistent naming convention. Some differences in file structure may be present, given the experimental nature of OPM files.

- opm/opm_us_ecf_sein_aux.sas7bdat
- opm/opm_us_ecf_sein.sas7bdat
- opm/opm_us_ecf_seinunit_aux.sas7bdat
- opm/opm_us_ecf_seinunit.sas7bdat
- opm/opm_us_ehf_controltotals.sas7bdat
- opm/opm_us_ehf_phf.sas7bdat
- opm/opm_us_ehf.sas7bdat
- opm/opm_us_ehf_shf.sas7bdat
- opm/opm_us_ehf_uhf.sas7bdat
- opm/opm_us_ehf_uniqpik.sas7bdat
- opm/opm_us_icf_aux.sas7bdat
- opm/opm_us_icf.sas7bdat
- opm/opm_us_jhf.sas7bdat
- opm/opm_us_qwi_seinunit_rh.sas7bdat
- opm/opm_us_qwi_seinunit_sa.sas7bdat
- opm/opm_us_qwi_seinunit_se.sas7bdat
- opm/opm_us_spf.sas7bdat
- opm/opm_us_u2w.sas7bdat

The following file is available with T26 permissions:

- opm/opm_us_ecf_t26.sas7bdat

### 6.3.2 Data location

The core files are stored in a `opm` sub-directories of the process-specific directories, similar to how state-specific files are stored

```
/opm/
```

Some California-sourced QCEW data from the OPM-ECF files are covered under Title 26, and can be found in

```
/opmt26/
```

### 6.3.3 Available processes

Generally, OPM records are identified by the same variables (PIK, SEIN, SEINUNIT) as UI-wage-record derived records. SEIN and SEINUNIT have been attached to the records by a name-and-address match to QCEW files. A mapping of SEIN to agency identifiers is available on demand.

#### 6.3.3.1 ECF

The OPM-ECF is processed the same way as the regular ECF, based on "pseudo-UI" records and matched QCEW records. Only establishments for federal agencies are in-scope to have a record in the OPM-ECF. The current version of OPM does not, however, include the firm-age and firm-size (or agency-age and agency-size) variables that were computed for the ECF. OPM entities that link to California records on the QCEW are available under Title 26 permissions, the same way as for the regular ECF. However, the current OPM processing uses an older version of ECF processing. Its file structure (variable names) are documented in Section 6.3.5 - minor differences between the files here and the files in chapter 2 are left to the user to resolve.

#### 6.3.3.2 EHF

The OPM-EHF is designed to store the complete federal government work history, for each individual that appears in the OPM source records, subject to the coverage limitations of OPM (see Table 6.8). The data has been restructured to resemble UI wage records. This involves some coarsening of the data (OPM source records contain the exact termination date of a job, whereas UI wage records do not). Otherwise, the structure of all EHF-like files is identical to those provided by the EHF process, and are documented in chapter 3 (this includes the JHF). The establishment identifier corresponds to the SEIN-SEINUNIT found when matching to the QCEW.

#### 6.3.3.3 ICF

While the structure of the OPM-ICF is identical to the ICF documented in chapter 5, none of the contents are derived from the data sources noted there. Rather, all demographic information is derived entirely from the OPM-provided input files. Section 6.1.3.8 describes how any missing data is filled in, and how race in particular is standardized across the two different coding schemata present in the OPM data. Due to the low incidence of missingness in the data, only a single implicate was generated for any imputes; thus, no implicate files are provided. In addition to the standard ICF, an auxiliary file with variables not usually present on the ICF is provided, see Section 6.3.4.1. Address data is not available.

#### 6.3.3.4 QWI-SEINUNIT files

Establishment-level files from the QWI series are provided, and are structured as QWI-SEINUNIT files were structured in S2011. While the description of these files in chapter 7 is mostly correct, users should consult Vilhuber and McKinney (2014) for the appropriate details. Note that the SEIN-SEINUNIT identifier is used, not the OPM agency identifier.

**Record identifier** YEAR QUARTER SEIN SEINUNIT
**Sort order** YEAR QUARTER SEIN SEINUNIT
**Entity** 'establishment' (not agency)
**Unique Entity Key** SEIN SEINUNIT

#### 6.3.3.5 U2W

Because there is some mismatch between establishment-level reports in the QCEW and the equivalent agency-within-county or agency-within-city reporting in the duty station file provided by OPM, the Unit-to-Worker Impute process is used to allocated to QCEW-reported establishments when a unique allocation is not feasible. The structure of the file is otherwise identical to the file(s) described in chapter 10.

### 6.3.4 Dataset documentation on unique files

#### 6.3.4.1 Auxiliary dataset for ICF: OPM_us_ICF_aux

This contains variables not otherwise present on the ICF, derived from OPM information. No effort has been made to fill in any missing information, the file is provided as-is to researchers.

**Record identifier** PIK
**Sort order** PIK
**Entity** PIK
**Unique Entity Key** PIK

Table 6.1: OPM_US_ICF_AUX: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **best_dob** | OPM Date of Birth (YYYYMM): bested | Char | 6 |
| **best_edlvl** | OPM edlvl: bested | Char | 2 |
| **best_erirace** | OPM (OMB-compliant) erirace: bested | Char | 6 |
| **best_race** | OPM (old-style) race: bested | Char | 1 |
| **doa** | Date of Arrival (SAS date value) | Num | 4 |
| **educ_f** | Highest grade achieved | Num | 3 |
| **fb** | Foreign-born status | Char | 1 |
| **flag_anyrace** | Either ERIRACE or RACE have a value | Num | 3 |
| **flag_best_dob** | BEST_DOB has a value | Num | 3 |
| **flag_best_edlvl** | BEST_EDLVL has a value | Num | 3 |
| **flag_best_erirace** | BEST_ERIRACE has a value | Num | 3 |
| **flag_best_race** | BEST_RACE has a value | Num | 3 |
| **flag_dob** | DOB has a value | Num | 8 |
| **flag_edlvl** | EDLVL has a value | Num | 8 |
| **flag_educ_c** | EDUC_C has a value | Num | 3 |
| **flag_erirace** | ERIRACE has a value | Num | 8 |
| **flag_race** | RACE has a value | Num | 8 |
| **flag_race_ethnicity** | RACE_ETHNICITY has a value | Num | 3 |
| **flag_sex** | SEX has a value | Num | 3 |
| **missing_anyrace** | Either best_erirace or best_RACE have a value | Num | 8 |
| **missing_citizen** | CITIZEN is missing a value | Num | 8 |
| **missing_creditmilsrv** | CREDITMILSRV is missing a value | Num | 8 |
| **missing_degreeyr** | DEGREEYR is missing a value | Num | 8 |
| **missing_dob** | DOB is missing a value | Num | 8 |
| **missing_edlvl** | EDLVL is missing a value | Num | 8 |

(cont.)

**Table 6.1 (cont.): OPM_US_ICF_AUX: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **missing_erirace** | ERIRACE is missing a value | Num | 8 |
| **missing_frozenserv** | FROZENSERV is missing a value | Num | 8 |
| **missing_handicap** | HANDICAP is missing a value | Num | 8 |
| **missing_occupation** | OCCUPATION is missing a value | Num | 8 |
| **missing_patco** | PATCO is missing a value | Num | 8 |
| **missing_race** | RACE is missing a value | Num | 8 |
| **missing_sex** | SEX is missing a value | Num | 8 |
| **missing_vstatus** | VSTATUS is missing a value | Num | 8 |
| **opm_citizen** | US Citizenship | Char | 1 |
| **opm_creditmilsrv** | Creditable Military Service | Char | 4 |
| **opm_degreeyr** | Yr Degree Cert Attained | Char | 4 |
| **opm_dob** | Date of Birth | Char | 6 |
| **opm_edlvl** | Education Level | Char | 2 |
| **opm_erirace** | ERI (Ethnicity/Race Indicator) | Char | 6 |
| **opm_frozenserv** | Frozen Service | Char | 4 |
| **opm_handicap** | Handicap | Char | 2 |
| **opm_occupation** | Occupation | Char | 4 |
| **opm_patco** | Occupational Category(PATCO) | Char | 1 |
| **opm_race** | Race or National Orgin | Char | 1 |
| **opm_sex** | sex | Char | 1 |
| **opm_src** | Source of OPM record (qtime) run | Num | 8 |
| **opm_vstatus** | Veterans Status | Char | 1 |
| **pcf_race** | PCF-compliant (not sourced) race information | Char | 1 |
| **pik** | PIK | Char | 9 |

### 6.3.5 Dataset documentation for OPM-ECF

#### 6.3.5.1 Main ECF SEINUNIT dataset: opm_us_ecf_seinunit

ECF SEINUNIT-level file, research variables only. The standard version of this file is documented in Section 2.3.3.

**Record identifier:** SEIN SEINUNIT YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER SEINUNIT
**File indexes:** none
**Entity** "establishment" or SESA
**Unique Entity Key** SEIN SEINUNIT

Table 6.2: OPM_US_ECF_SEINUNIT: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **ES_COUNTY** | Cleaned ES202 FIPS County CCC | Char | 3 |
| **ES_EIN** | Cleaned EIN | Char | 9 |
| **ES_NAICS_FNL1997** | Final 1997 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_FNL2002** | Final 2002 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_FNL2007** | Final 2007 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_FNL1997_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_FNL2002_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_FNL2007_SRC** | Source of Ind Code | Char | 3 |
| **ES_OWNER_CODE** | Cleaned OWNER_CODE O | Char | 1 |
| **ES_SIC** | Cleaned SIC Code IIII | Char | 4 |
| **ES_SIC_DIV** | Cleaned SIC Division I | Char | 1 |
| **ES_SIC_SRC** | Source of Ind Code | Char | 3 |
| **NUM_ESTABS** | Number of Establishments | Num | 4 |
| **best_emp1** | Best UI/202 Employment Month 1 | Num | 4 |
| **best_emp2** | Best UI/202 Employment Month 2 | Num | 4 |
| **best_emp3** | Best UI/202 Employment Month 3 | Num | 4 |
| **best_flag** | Source of best_ data | Num | 3 |
| **best_wages** | Best UI/202 Wages | Num | 5 |
| **es_county_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_ein_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_fnl1997_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_fnl2002_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_fnl2007_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_owner_code_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_sic_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |

(cont.)

**Table 6.2 (cont.): OPM_US_ECF_SEINUNIT: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **es_state** | ES202 FIPS State SS | Char | 2 |
| **leg_block** | Census Block | Char | 4 |
| **leg_block_suf1** | Census Block suffix 1 | Char | 1 |
| **leg_block_suf2** | Census Block suffix 2 | Char | 1 |
| **leg_cbsa** | Core-Based Statistical Area | Char | 5 |
| **leg_cbsa_memi** | CBSA type 1=Metro, 2=Micro, else=9 | Char | 1 |
| **leg_county** | Cleaned GEO FIPS County CCC | Char | 3 |
| **leg_county_orig** | Cleaned GEO FIPS County CCC, pre-longitudinal impute | Char | 3 |
| **leg_galid** | Final GALID | Char | 29 |
| **leg_galid_orig** | GALID, pre-longitudinal impute | Char | 29 |
| **leg_geo_qual** | Quality of final geography | Num | 3 |
| **leg_geo_qual_orig** | Quality of geography, pre-longitudinal impute | Num | 3 |
| **leg_geocode** | FIPS Tab State\|\|FIPS Tab County\|\|Census Tract | Char | 11 |
| **leg_latitude** | Latitude, 6 implied decimal places | Num | 8 |
| **leg_longitude** | Longitude, 6 implied decimal places | Num | 8 |
| **leg_state** | Cleaned GEO State SS | Char | 2 |
| **leg_subctygeo** | Sub-county Geography from the LEG | Char | 10 |
| **leg_wib** | Workforce Investment Board area | Char | 6 |
| **multi_unit** | SEIN w/2+ records on 202 | Num | 3 |
| **qcew_auxiliary_code** | Firm engaged (not) engaged in production | Char | 1 |
| **quarter** | Quarter (numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **source** | 1=Earnings data only,2=202 only,3=both | Num | 3 |
| **ui_ein** | | Num | 8 |
| **ui_ein_miss** | | Num | 8 |
| **year** | Year YYYY | Num | 3 |
| **yr_qtr** | Continuous Time YEAR QUARTER | Char | 6 |

### 6.3.6 Auxiliary SEINUNIT dataset: opm_us_ecf_seinunit_aux

ECF SEINUNIT-level file, auxiliary and diagnostic variables only. The standard version of this file is documented in Section 2.3.4.

**Record identifier:** SEIN SEINUNIT YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER SEINUNIT

**File indexes:** none
**Entity** “establishment” or SESA
**Unique Entity Key** SEIN SEINUNIT

Table 6.3: OPM_US_ECF_SEINUNIT_AUX: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **ES_NAICS1997** | Cleaned 1997 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS2002** | Cleaned 2002 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS2007** | Cleaned 2007 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS1997_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS2002_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS2007_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_AUX1997** | Cleaned 1997 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_AUX2002** | Cleaned 2002 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_AUX2007** | Cleaned 2007 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_AUX1997_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_AUX2002_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_AUX2007_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_ESO1997** | ES202 ONLY 1997 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_ESO2002** | ES202 ONLY 2002 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_ESO2007** | ES202 ONLY 2007 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_ESO1997_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_ESO2002_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_ESO2007_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_IMP1997** | SIC IMP 1997 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_IMP2002** | SIC IMP 2002 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_IMP2007** | SIC IMP 2007 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_IMP1997_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_IMP2002_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_IMP2007_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_LDB1997** | Cleaned 1997 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_LDB2002** | Cleaned 2002 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_LDB2007** | Cleaned 2007 NAICS Code NNNNNN | Char | 6 |
| **ES_NAICS_LDB1997_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_LDB2002_SRC** | Source of Ind Code | Char | 3 |
| **ES_NAICS_LDB2007_SRC** | Source of Ind Code | Char | 3 |
| **county_impute_source** | 1=county 2=es_county (long edit) 3=mode_leg_county_emp 4=mode_leg_county 5=leg_county_orig | Num | 3 |

(cont.

**Table 6.3 (cont.): OPM_US_ECF_SEINUNIT_AUX: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **es_county_flag** | Quarters Away County data found | Num | 3 |
| **es_ein_flag** | Quarters Away EIN data found | Num | 3 |
| **es_galid** | GALID of address on es202 | Char | 29 |
| **es_naics1997_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics1997_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics1997_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics2002_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics2002_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics2002_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics2007_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics2007_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics2007_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics_aux1997_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics_aux1997_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_aux1997_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics_aux2002_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics_aux2002_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_aux2002_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics_aux2007_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics_aux2007_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_aux2007_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics_eso1997_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_eso2002_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_eso2007_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_imp1997_miss** | 0=ok,1=not found,2+found off qtr | Num | 8 |
| **es_naics_imp2002_miss** | 0=ok,1=not found,2+found off qtr | Num | 8 |
| **es_naics_imp2007_miss** | 0=ok,1=not found,2+found off qtr | Num | 8 |
| **es_naics_ldb1997_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics_ldb1997_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_ldb1997_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics_ldb2002_flag** | Quarters Away NAICS data found | Num | 3 |
| **es_naics_ldb2002_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_ldb2002_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_naics_ldb2007_flag** | Quarters Away NAICS data found | Num | 3 |

(cont.)

**Table 6.3 (cont.): OPM_US_ECF_SEINUNIT_AUX: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **es_naics_ldb2007_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **es_naics_ldb2007_valid** | Seinunit has some NAICS info | Num | 3 |
| **es_owner_code_flag** | Quarters Away OWNER_CODE data found | Num | 3 |
| **es_sic_flag** | Quarters Away SIC data found | Num | 3 |
| **es_sic_valid** | Seinunit has some SIC info | Num | 3 |
| **geo_vars_flag** | Quarters Away LEG variables found | Num | 3 |
| **geo_vars_miss** | 0=ok,1=not found,2+found off qtr label | Num | 3 |
| **leg_flag_geo** | Flag, number of quarters to find geocodes | Num | 3 |
| **naics_1997_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_2002_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_2007_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_aux_1997_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_aux_2002_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_aux_2007_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_ldb_1997_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_ldb_2002_invalid** | NAICS Code not Valid | Char | 1 |
| **naics_ldb_2007_invalid** | NAICS Code not Valid | Char | 1 |
| **qcew_county** | Original ES202 County | Char | 3 |
| **qcew_ein** | Original ES202 EIN | Char | 9 |
| **qcew_ein_bad** | Letters a-z,A-Z in EIN | Num | 3 |
| **qcew_ein_defect** | Problem with EIN | Num | 3 |
| **qcew_empl_month1** | Original ES202 Employment Month 1 | Num | 5 |
| **qcew_empl_month2** | Original ES202 Employment Month 2 | Num | 5 |
| **qcew_empl_month3** | Original ES202 Employment Month 3 | Num | 5 |
| **qcew_empl_month1_flg** | Reported or imputed Month 1 Employment | Char | 1 |
| **qcew_empl_month2_flg** | Reported or imputed Month 2 Employment | Char | 1 |
| **qcew_empl_month3_flg** | Reported or imputed Month 3 Employment | Char | 1 |
| **qcew_naics1997** | Original NAICS 1997 Code | Char | 6 |
| **qcew_naics2002** | Original NAICS 2002 Code | Char | 6 |
| **qcew_naics2007** | Original NAICS 2007 Code | Char | 6 |
| **qcew_naics_aux1997** | Original NAICS AUX 1997 Code | Char | 6 |
| **qcew_naics_aux2002** | Original NAICS AUX 2002 Code | Char | 6 |
| **qcew_naics_aux2007** | Original NAICS AUX 2007 Code | Char | 6 |
| **qcew_naics_ldb1997** | Original NAICS LDB 1997 Code | Char | 6 |

(cont.

**Table 6.3 (cont.): OPM_US_ECF_SEINUNIT_AUX: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **qcew_naics_ldb2002** | Original NAICS LDB 2002 Code | Char | 6 |
| **qcew_naics_ldb2007** | Original NAICS LDB 2007 Code | Char | 6 |
| **qcew_owner_code** | Original Owner Code | Char | 1 |
| **qcew_sic** | Original ES202 SIC | Char | 4 |
| **qcew_sic_invalid** | SIC Code not Valid | Char | 1 |
| **qcew_total_wages** | Original ES202 wages | Num | 5 |
| **qcew_total_wages_flg** | Reported or imputed Total Wages | Char | 1 |
| **qcew_valid_ein** | EIN in known IRD | Num | 3 |
| **quarter** | Quarter (numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **seinunit_bad** | SEINUNIT data non-numeric | Num | 3 |
| **seinunit_type** | 0 if seinunit=00000 | Num | 3 |
| **special_handle** | candidate for structure fix | Num | 3 |
| **structure_fix** | Multiunit Imputed Record Structure | Num | 3 |
| **ui_ein_flag** | | Num | 8 |
| **year** | Year YYYY | Num | 3 |

### 6.3.7 Main SEIN dataset: opm_us_ecf_sein

ECF SEIN-level file, with variables aggregated from the establishment level. The standard version of this file is documented in Section 2.3.5.

**Record identifier:** SEIN YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER
**File indexes:** none
**Entity** "firm"
**Unique Entity Key** SEIN

Table 6.4: OPM_US_ECF_SEIN: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **MODE_ES_COUNTY** | Unit Mode Cleaned County | Char | 3 |
| **MODE_ES_COUNTY_EMP** | Emp Mode Cleaned County | Char | 3 |
| **MODE_ES_EIN** | Unit Mode Cleaned EIN | Char | 9 |
| **MODE_ES_EIN_EMP** | Emp Mode Cleaned EIN | Char | 9 |
| **MODE_ES_NAICS_FNL1997** | Unit Mode Cleaned NAICS 1997 | Char | 6 |
| **MODE_ES_NAICS_FNL2002** | Unit Mode Cleaned NAICS 2002 | Char | 6 |
| **MODE_ES_NAICS_FNL2007** | Unit Mode Cleaned NAICS 2007 | Char | 6 |
| **MODE_ES_NAICS_FNL1997_EMP** | Emp Mode Cleaned NAICS 1997 | Char | 6 |
| **MODE_ES_NAICS_FNL2002_EMP** | Emp Mode Cleaned NAICS 2002 | Char | 6 |
| **MODE_ES_NAICS_FNL2007_EMP** | Emp Mode Cleaned NAICS 2007 | Char | 6 |
| **MODE_ES_OWNER_CODE** | Unit Mode Cleaned OWNER_CODE | Char | 1 |
| **MODE_ES_OWNER_CODE_EMP** | Emp Mode Cleaned OWNER_CODE | Char | 1 |
| **MODE_ES_SIC** | Unit Mode Cleaned SIC | Char | 4 |
| **MODE_ES_SIC_EMP** | Emp Mode Cleaned SIC | Char | 4 |
| **MODE_LEG_CBSA** | Unit Mode Cleaned GEO CBSA | Char | 5 |
| **MODE_LEG_CBSA_MEMI** | Unit Mode Cleaned GEO CBSA type | Char | 1 |
| **MODE_LEG_COUNTY** | Unit Mode Cleaned GEO COUNTY | Char | 3 |
| **MODE_LEG_COUNTY_EMP** | Emp Mode Cleaned GEO COUNTY | Char | 3 |
| **MODE_LEG_COUNTY_ORIG** | Unit Mode Cleaned GEO COUNTY, pre-longitudinal impute | Char | 3 |
| **MODE_LEG_COUNTY_ORIG_EMP** | Emp Mode Cleaned GEO COUNTY, pre-longitudinal impute | Char | 3 |
| **MODE_LEG_STATE** | Unit Mode Cleaned GEO STATE | Char | 2 |
| **MODE_LEG_STATE_EMP** | Emp Mode Cleaned GEO STATE | Char | 2 |
| **MODE_LEG_SUBCTYGEO** | Unit Mode Cleaned GEO COUNTY | Char | 10 |
| **MODE_LEG_SUBCTYGEO_EMP** | Emp Mode Cleaned GEO COUNTY | Char | 10 |
| **MODE_LEG_WIB** | Unit Mode Cleaned GEO WIB | Char | 6 |
| **MODE_LEG_WIB_EMP** | Emp Mode Cleaned GEO WIB | Char | 6 |

(cont.)

Table 6.4 (cont.): OPM_US_ECF_SEIN: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **NUM_ESTABS** | Number of Establishments | Num | 4 |
| **mode_es_county_emp_miss** | Missing Value | Num | 3 |
| **mode_es_county_miss** | Missing Value | Num | 3 |
| **mode_es_ein_emp_miss** | Missing Value | Num | 3 |
| **mode_es_ein_miss** | Missing Value | Num | 3 |
| **mode_es_naics_fnl1997_emp_miss** | Missing Value | Num | 3 |
| **mode_es_naics_fnl1997_miss** | Missing Value | Num | 3 |
| **mode_es_naics_fnl2002_emp_miss** | Missing Value | Num | 3 |
| **mode_es_naics_fnl2002_miss** | Missing Value | Num | 3 |
| **mode_es_naics_fnl2007_emp_miss** | Missing Value | Num | 3 |
| **mode_es_naics_fnl2007_miss** | Missing Value | Num | 3 |
| **mode_es_owner_code_emp_miss** | Missing Value | Num | 3 |
| **mode_es_owner_code_miss** | Missing Value | Num | 3 |
| **mode_es_sic_emp_miss** | Missing Value | Num | 3 |
| **mode_es_sic_miss** | Missing Value | Num | 3 |
| **mode_leg_cbsa_emp** | Emp Mode Cleaned GEO CBSA | Char | 5 |
| **mode_leg_cbsa_emp_miss** | Missing Value | Num | 3 |
| **mode_leg_cbsa_memi_emp** | Emp Mode Cleaned GEO CBSA type | Char | 1 |
| **mode_leg_cbsa_miss** | Missing Value | Num | 3 |
| **mode_leg_county_emp_miss** | Missing Value | Num | 3 |
| **mode_leg_county_miss** | Missing Value | Num | 3 |
| **mode_leg_county_orig_emp_miss** | Missing Value | Num | 8 |
| **mode_leg_county_orig_miss** | Missing Value | Num | 8 |
| **mode_leg_state_emp_miss** | Missing Value | Num | 3 |
| **mode_leg_state_miss** | Missing Value | Num | 3 |
| **mode_leg_subctygeo_emp_miss** | Missing Value | Num | 3 |
| **mode_leg_subctygeo_miss** | Missing Value | Num | 3 |
| **mode_leg_wib_emp_miss** | Missing Value | Num | 3 |
| **mode_leg_wib_miss** | Missing Value | Num | 3 |
| **multi_first_quarter** | First Quarter SEIN on 202 | Num | 3 |
| **multi_first_year** | First Year SEIN on 202 | Num | 3 |
| **multi_unit** | SEIN w/2+ records on 202 | Num | 3 |
| **multi_unit_code** | ES202 multi-unit (non) reporter | Char | 1 |
| **quarter** | Quarter (numeric) | Num | 3 |

(cont.)

**Table 6.4 (cont.): OPM_US_ECF_SEIN: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **qwi_unit_weight** | Weight sum(B_UI)=sum(month1_BLS) | Num | 8 |
| **sein** | State Employer Identification Number | Char | 12 |
| **sein_best_emp1** | SEIN Best UI/202 Month 1, Employment | Num | 4 |
| **sein_best_emp2** | SEIN Best UI/202 Month 2, Employment | Num | 4 |
| **sein_best_emp3** | SEIN Best UI/202 Month 3, Employment | Num | 4 |
| **sein_best_wages** | SEIN Best UI/202 Payroll | Num | 5 |
| **source** | 1=Earnings data only,2=202 only,3=both | Num | 3 |
| **year** | Year YYYY | Num | 3 |

### 6.3.8 Auxiliary SEIN dataset: opm_us_ecf_sein_aux

ECF SEIN-level file, auxiliary and diagnostic variables only. The standard version of this file is documented in Section 2.3.6.

**Record identifier:** SEIN YEAR QUARTER
**Sort order:** SEIN YEAR QUARTER
**File indexes:** none
**Entity** “firm”
**Unique Entity Key** SEIN

Table 6.5: OPM_US_ECF_SEIN_AUX: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **MODE_ES_NAICS_ESO1997** | Unit Mode Cleaned NAICS 1997 | Char | 6 |
| **MODE_ES_NAICS_ESO2002** | Unit Mode Cleaned NAICS 2002 | Char | 6 |
| **MODE_ES_NAICS_ESO2007** | Unit Mode Cleaned NAICS 2007 | Char | 6 |
| **MODE_ES_NAICS_ESO1997_EMP** | Emp Mode Cleaned NAICS 1997 | Char | 6 |
| **MODE_ES_NAICS_ESO2002_EMP** | Emp Mode Cleaned NAICS 2002 | Char | 6 |
| **MODE_ES_NAICS_ESO2007_EMP** | Emp Mode Cleaned NAICS 2007 | Char | 6 |
| **emp1_UI** | Best SEIN UI Employment | Num | 4 |
| **emp2_UI** | Best SEIN UI Employment | Num | 4 |
| **emp3_UI** | Best SEIN UI Employment | Num | 4 |
| **ever_202** | SEIN ever on 202 | Num | 3 |
| **ever_UI** | SEIN ever on UI | Num | 3 |
| **ever_emp1** | MULTI ever has ES202 month 1 employment | Num | 3 |
| **ever_emp2** | MULTI ever has ES202 month 2 employment | Num | 3 |
| **ever_emp3** | MULTI ever has ES202 month 3 employment | Num | 3 |
| **ever_multi** | SEIN ever multi unit | Num | 3 |
| **ever_wages** | MULTI ever ES202 wages | Num | 3 |
| **in_202** | SEIN in ES202 | Num | 3 |
| **in_UI** | SEIN in UI | Num | 3 |
| **master_empl_month1_flg** | Stored Master Record Flag | Char | 1 |
| **master_empl_month2_flg** | Stored Master Record Flag | Char | 1 |
| **master_empl_month3_flg** | Stored Master Record Flag | Char | 1 |
| **master_multi_unit_code** | Stored Master Multi Code | Char | 1 |
| **master_total_wages_flg** | Stored Master Record Flag | Char | 1 |
| **mode_es_county_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_county_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_ein_emp_flag** | Quarters Away Data Found | Num | 3 |

(cont.)

Table 6.5 (cont.): OPM_US_ECF_SEIN_AUX: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **mode_es_ein_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_eso1997_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_eso1997_emp_miss** | Missing Value | Num | 3 |
| **mode_es_naics_eso1997_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_eso1997_miss** | Missing Value | Num | 3 |
| **mode_es_naics_eso2002_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_eso2002_emp_miss** | Missing Value | Num | 3 |
| **mode_es_naics_eso2002_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_eso2002_miss** | Missing Value | Num | 3 |
| **mode_es_naics_eso2007_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_eso2007_emp_miss** | Missing Value | Num | 3 |
| **mode_es_naics_eso2007_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_eso2007_miss** | Missing Value | Num | 3 |
| **mode_es_naics_fnl1997_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_fnl1997_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_fnl2002_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_fnl2002_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_fnl2007_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_naics_fnl2007_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_owner_code_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_owner_code_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_sic_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_sic_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_cbsa_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_cbsa_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_county_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_county_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_county_orig_emp_flag** | Quarters Away Data Found | Num | 8 |
| **mode_leg_county_orig_flag** | Quarters Away Data Found | Num | 8 |
| **mode_leg_state_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_state_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_subctygeo_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_subctygeo_flag** | Quarters Away Data Found | Num | 3 |
| **mode_leg_wib_emp_flag** | Quarters Away Data Found | Num | 3 |

(cont.)

**Table 6.5 (cont.): OPM_US_ECF_SEIN_AUX: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **mode_leg_wib_flag** | Quarters Away Data Found | Num | 3 |
| **qcew_sein_emp1** | SEIN 202 Employment Month 1 | Num | 4 |
| **qcew_sein_emp2** | SEIN 202 Employment Month 2 | Num | 4 |
| **qcew_sein_emp3** | SEIN 202 Employment Month 3 | Num | 4 |
| **qcew_sein_wages** | SEIN 202 Wages | Num | 5 |
| **quarter** | Quarter (numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **ui_payroll** | Original UI Payroll Info W1 | Num | 5 |
| **ui_seinsize_b** | UI Employment B | Num | 4 |
| **ui_seinsize_e** | UI Employment E | Num | 4 |
| **ui_seinsize_m** | UI Employment M | Num | 4 |
| **ui_wages** | SEIN UI Wages | Num | 5 |
| **year** | Year YYYY | Num | 3 |

### 6.3.9 Auxiliary SEIN dataset: opm_us_ecf_t26

ECF T26 variables associated with both the SEIN and the SEINUNIT-level file. For California, this includes the EIN-related variables as well. For all states, this includes any variables derived from T26 datasets, primarily the BR. The standard version of this file is documented in Section 2.3.8.

**Record identifier:** SEIN SEINUNIT YEAR QUARTER
**Sort order:** SEIN SEINUNIT YEAR QUARTER
**File indexes:** none
**Entity** "establishment" or SESA
**Unique Entity Key** SEIN SEINUNIT

Table 6.6: OPM_US_ECF_T26: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **ES_EIN** | Cleaned EIN | Char | 9 |
| **MODE_ES_EIN** | Unit Mode Cleaned EIN | Char | 9 |
| **MODE_ES_EIN_EMP** | Emp Mode Cleaned EIN | Char | 9 |
| **es_ein_flag** | Quarters Away EIN data found | Num | 3 |
| **es_ein_miss** | 0=ok,1=not found,2+found off qtr | Num | 3 |
| **mode_es_ein_emp_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_ein_emp_miss** | Missing Value | Num | 3 |
| **mode_es_ein_flag** | Quarters Away Data Found | Num | 3 |
| **mode_es_ein_miss** | Missing Value | Num | 3 |
| **qcew_ein** | Original ES202 EIN | Char | 9 |
| **qcew_ein_bad** | Letters a-z,A-Z in EIN | Num | 3 |
| **qcew_ein_defect** | Problem with EIN | Num | 3 |
| **qcew_valid_ein** | EIN in known IRD | Num | 3 |
| **quarter** | Quarter (numeric) | Num | 3 |
| **sein** | State Employer Identification Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **ui_ein** | | Num | 8 |
| **ui_ein_flag** | | Num | 8 |
| **ui_ein_miss** | | Num | 8 |
| **year** | Year YYYY | Num | 3 |

## 6.4 NOTES

## 6.5 TABLES

Table 6.7: Non-reporting agencies in OPM

| **Notable non-reporting agencies to OPM** |
|---|
| *Some Executive Branch agencies:*<br>• Office of the President and of the Vice President<br>• White House<br>• Office of Policy Development<br>• multiple Department of Defense agencies, including military personnel<br>• most Intelligence agencies<br>• Board of Governors of the Federal Reserve<br>• Tennessee Valley Authority<br>• State Department Regional Personnel Centers<br>• U.S. Postal Service<br>• Other smaller federal agencies or commissions. |
| *Most of the Legislative Branch, such as:*<br>• Congress<br>• Congressional Budget Office<br>• Government Accountability Office<br>• Library of Congress<br>• Office of Compliance |
| *Judicial branch completely, including*<br>• Supreme Court<br>• U.S. Courts |

Table 6.8: Exclusions from LEHD federal worker universe

| These agencies are excluded because of suppressed geography in public-use data: | |
|---|---|
| DJ02 | FEDERAL BUREAU OF INVESTIGATION |
| DJ06 | DRUG ENFORCEMENT ADMINISTRATION |
| DJ15 | BUREAU OF ALCOHOL, TOBACCO, FIREARMS, AND EXPLOSIVES (ATF) |
| HSAD | U.S. SECRET SERVICE |
| TR40 | ALCOHOL AND TOBACCO TAX AND TRADE BUREAU |
| TRAC | U.S. SECRET SERVICE |
| TRAD | U.S. MINT |

Table 6.9: OPM: Employment in agencies that do not report geography

| Agency | DC area | Other US |
|---|---:|---:|
| DJ02 | 10468 | 10468 |
| DJ02 | 23018 | 23018 |
| DJ06 | 2269 | 2269 |
| DJ06 | 7354 | 7354 |
| DJ15 | 1118 | 1118 |
| DJ15 | 4001 | 4001 |
| HSAD | 3948 | 3948 |
| HSAD | 2854 | 2854 |
| TR40 | 159 | 159 |
| TR40 | 357 | 357 |
| TRAD | 364 | 364 |
| TRAD | 1429 | 1429 |

Note: Data from public-use OPM, 2009Q4.

Table 6.10: OPM-QCEW Matching stragegy

| *Match pass* | *Match type* | *Match description* | *Blocking vars* | *Matching vars* |
|---|---|---|---|---|
| Automated matches | | | | |
| 1 | DQ95F5 | Automated Matches on agency names within dept/ county, fuzzy employment | state county dept_full | agency(95)<br>aggemp_post(95) |
| 2 | DQ85F5 | Automated Matches on agency names within dept/ county, fuzzy employment | state county dept_full | agency(85)<br>aggemp_post(95) |
| 3 | DQ95F15 | Automated Matches on agency names within dept/ county, fuzzy employment | state county dept_full | agency(95)<br>aggemp_post(85) |
| 4 | DQ85F15 | Automated Matches on agency names within dept/ county, fuzzy employment | state county dept_full | agency(85)<br>aggemp_post(85) |
| 5 | DQ95F25 | Automated Matches on agency names within dept/ county, fuzzy employment | state county dept_full | agency(95)<br>aggemp_post(75) |
| 6 | DQ95 | Automated Matches on agency names within dept/ county, no employment | state county dept_full | agency(95) |
| 7 | UNIQ1 | Uniques Finds agencies that are the only agency for DEPT within COUNTY | DEPT COUNTY count(agency)=1 | n.a. |
| Matches subject to clerical review | | | | |
| 61 | SEIN F5 | PreClerical Find SEINUNITs with assumed SEIN from x-county editing, ignore agency name | state county dept_full sein | aggemp_post(95) |
| 62 | SEIN F15 | PreClerical Find SEINUNITs with assumed SEIN from x-county editing, ignore agency name | state county dept_full sein | aggemp_post(85) |
| 63 | DQ95 | PreClerical Relax department, match on agency name alone | state county | agency(95) |
| 64 | DQ95F5 | PreClerical Ignore agency name, use fuzzy employment and dept | state county dept_full | aggemp_post(95) |
| 65 | DQ95F15 | PreClerical Ignore agency name, use fuzzy employment and dept | state county dept_full | aggemp_post(85) |
| 66 | DQ95F25 | PreClerical Ignore agency name, use fuzzy employment and dept | state county dept_full | aggemp_post(85) |
| 67 | DQ95F5 | PreClerical Ignore county, use fuzzy employment and agency | state dept_full | agency(95)<br>aggemp_post(95) |
| 68 | DQ95F15 | PreClerical Ignore county, use fuzzy employment and agency | state dept_full | agency(95)<br>aggemp_post(85) |
| 71-78 | | PreClerical Repeat above sequence, but allow for re-matches | see above | see above |
| SEIN Imputes | | | | |
| 81 | IMP | PreImpute Impute SEIN(not SEINUNIT) based on conditional distributions These units will need to go through U2W!! | state county dept_full sizeclass | random(employment distribution) |
| 82 | IMP | PreImpute Impute SEIN(not SEINUNIT) based on conditional distributions These units will need to go through U2W!! | state county dept_full | random(employment distribution) |
| 83 | IMP | PreImpute Impute SEIN(not SEINUNIT) based on conditional distributions These units will need to go through U2W!! | state dept_full sizeclass | random(employment distribution) |
| 84 | IMP | PreImpute Impute SEIN(not SEINUNIT) based on conditional distributions These units will need to go through U2W!! | state dept_full | random(employment distribution) |

Table 6.10 (cont.) OPM-QCEW Matching stragegy

| *Match* | *Match* | | | |
|---|---|---|---|---|
| *pass* | *type* | *Match description* | *Blocking vars* | *Matching vars* |
| 85 | IMP | PreImpute Impute SEIN(not SEINUNIT) based on conditional distributions These units will need to go through U2W!! | state county | random(employment distribution) |
| 86 | IMP | PreImpute Impute SEIN(not SEINUNIT) based on conditional distributions These units will need to go through U2W!! | state | random(employment distribution) |
| 91-96 | | PreImpute Repeat above sequence, but allow for re-matches | | |

Table 6.11: OPM: DHS Reorganization 2003

| Original Agency | Original Department | Current Agency or Office (in DHS) |
|---|---|---|
| U.S. Customs Service | Treasury | U.S. Customs and Border Protection |
| | | U.S. Immigration and Customs Enforcement |
| Immigration and Naturalization Service | Justice | U.S. Customs and Border Protection |
| | | U.S. Immigration and Customs Enforcement |
| | | U.S. Citizenship and Immigration Services |
| Federal Protective Service | General Services Administration (GSA) | National Protection and Programs Directorate |
| Transportation Security Administration | Transportation | Transportation Security Administration |
| Federal Law Enforcement Training Center | Treasury | Federal Law Enforcement Training Center |
| Animal and Plant Health Inspection Service (part) | Agriculture | U.S. Customs and Border Protection |
| Federal Emergency Management Agency (FEMA) | none | Federal Emergency Management Agency |
| Office for Domestic Preparedness | Justice | Responsibilities distributed within FEMA |
| Strategic National Stockpile, National Disaster Medical System | Health and Human Services (HHS) | Returned to HHS, July, 2004 |
| Nuclear Incident Response Team | Energy | Responsibilities distributed within FEMA |
| Domestic Emergency Support Teams | Justice | |
| National Domestic Preparedness Office | FBI | |
| CBRN Countermeasures Programs | Energy | Science & Technology Directorate |
| Environmental Measurements Laboratory | Energy | |
| National Biological Warfare, Defense Analysis Center | Defense | |
| Plum Island Animal Disease Center | Agriculture | |
| Federal Computer Incident Response Center | GSA | US-CERT, Office of Cybersecurity and Communications |
| | | National Protection and Programs Directorate |
| National Communications System | Defense | Office of Cybersecurity and Communications |
| | | National Protection and Programs Directorate |
| National Infrastructure Protection Center | FBI | Office of Operations Coordination |
| | | Office of Infrastructure Protection |
| Energy Security and Assurance Program | Energy | Office of Infrastructure Protection |
| U.S. Coast Guard | Transportation | U.S. Coast Guard |
| U.S. Secret Service | Treasury | U.S. Secret Service |

Source: https://en.wikipedia.org/wiki/United_States_Department_of_Homeland_Security, accessed 2012-04-17, and http://www.dhs.gov/xabout/history/editorial_0133.shtm

# Chapter 7. Quarterly Workforce Indicators - SEINUNIT file (QWI)

## 7.1 OVERVIEW

The Quarterly Workforce Indicators (QWI) establishment file contains quarterly measures of workforce composition and worker turnover at the establishment level. The LEHD establishment-level measures are created from longitudinally integrated person and establishment-level data. Establishment-level measures include: (i) Worker and Job Flows: accessions, separations, job creation, job destruction by age and gender of workforce; (ii) Worker composition by gender and age, (iii) Worker compensation for stocks and flows by gender and age; (iv) Dynamic worker compensation summary statistics for stocks and flows by gender and age. The QWI may be used in combination with the ECF (chapter 2) to match to other Census micro business databases, and can be linked by firm-establishment identifiers to other LEHD Infrastructure files.

### 7.1.1 Changes in this Snapshot

The QWI_SEINUNIT files (internally known as UFF_B) have been modified since the previous snapshot, reflecting changes in the QWI publications for which they serve as inputs. Each file contains the statistics known from the public-use QWI. Variable names have changed. The QWI_SEINUNIT files contain no FTI.

#### 7.1.1.1 Restrictions

Note that the use of the QWI_SEINUNIT files is incompatible with the use of the QWI public-use files also available in the FSRDC. Researchers must choose one or the other.

#### 7.1.1.2 Changes to names of files

Prior to S2011, only *age x sex* tabulations were available, and the files were simply called "QWI_SEINUNIT". In S2011, race, ethnicity, and education tabulations were added. In this release, two file names have been modified:

- QWI_SEINUNIT_SA is the new name of QWI_SEINUNIT_WIA, containing *age x sex* statistics
- QWI_SEINUNIT_D is the new name of QWI_SEINUNIT_estabtots, containing only the marginal categories (i.e., no breakouts by demographic specific groups)

In addition, the marginal categories contained in QWI_SEINUNIT_D are no longer duplicated in the other files.

#### 7.1.1.3 Renamed and dropped variables

In line with changed publication of QWI (lehd.ces.census.gov/doc/Memo_changes_to_QWI.pdf), several variables are no longer present, or have been renamed. In general, names are consistent with the "Alternate names" in LEHD Schema V4.0.1 (see lehd.ces.census.gov/data/schema/V4.0.1).

The following variables are no longer available:

- All variables related to *changes in total earnings* (starting with `dW`)
- All variables that are rates (ending in `R`)
- All variables related to periods of non-employment preceding or following a transition (starting with `N`)
- Certain average earnings variables (`WCA, WCS, WA, WS`)
- Certain variables relating to continuous quarter hiring (`CH`, `CR`)
- The variable `FSnx` (Full-quarter separations in the next quarter), due to a processing change (CHECK)
- The variable `W2` (average earnings for end-of-quarter employment), replaced by `W2B` (average earnings for beginning-of-quarter employment)

The following variables have been renamed:

- The name of the SIC variable has changed from `ES_SIC` to `SIC1987FNL`
- The name of the NAICS variable has changed from `ES_NAICS_FNL2007` to `NAICS2012fnl` and reflects NAICS 2012 coding.
- The name of `H3` (New Hires into Full-Quarter Employment) has been changed to `FH` for consistency with the public-use variables.
- Earnings-related variables have been made consistent with the public-use variables

| Public-use name | Internal name | Label |
|---|---|---|
| ZW3 | W3 | Average Monthly Earnings (Full-Quarter Employment) |
| ZW1 | W2B | Average Monthly Earnings (Beginning-of-Quarter Employment) |
| ZWFA | WFA | Average Monthly Earnings (All Hires into Full-Quarter Employment) |
| ZWFH | WH3 | Average Monthly Earnings (New Hires into Full-Quarter Employment) |
| ZWFS | WFS | Average Monthly Earnings (Flows out of Full-Quarter Employment) |

#### 7.1.1.4 Changes in coding (variable names)

Please note that coding for race and ethnicity has changed, see Table 7.1. This affects the naming of variables.

Table 7.1: QWI race and ethnicity coding in S2011 and S2014

| Label | S2011 | S2014 |
|---|---|---|
| | Ethnicity coding | |
| All (Any Ethinicity) | H0 | A0 |
| Hispanic or Latino | H1 | A2 |
| Not Hispanic or Latino | H2 | A1 |
| | Race coding | |
| All Races | R0 | A0 |
| American Indian or Alaska Native Alone | R1 | A3 |
| Asian Alone | R2 | A4 |
| Black or African American Alone | R3 | A2 |
| Native Hawaiian or Other Pacific Islander Alone | R4 | A5 |
| White Alone | R5 | A1 |
| Two or More Race Groups | R6 | A7 |

For S2011 coding, see download.vrdc.cornell.edu/qwipu/QWI-cheatsheet.txt.
For S2014 coding, see LEHD Schema V4.0.1 at lehd.ces.census.gov/data/schema/V4.0.1.

## 7.2 DATA CITATION

U.S. Census Bureau. 2016. *Quarterly Workforce Indicators (QWI) for establishments, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

## 7.3 DATA SET DESCRIPTIONS

### 7.3.1 Coverage of QWI

QWI data are available for all states that are LED-state partners, however, not every state is currently a LED-state partner. The QWI are built upon wage records in the UI system and information from state ES-202 data. The universe of QWI data is UI-covered earnings. UI coverage is broad, covering over 90% of total wage and salary civilian jobs.

When QWI private industry employment numbers are compared with other employment data, exclusions to UI coverage should be taken into account. Federal government employment is not generally included. Exempted employment varies slightly from state to state due to variations in state unemployment laws, but generally also excludes many farmers and agricultural employees, domestic workers, self-employed non-agricultural workers, members of the Armed Services, some state and local government employees as well as certain types of nonprofit employers and religious organizations (which are given a choice of coverage or noncoverage in a number of states). See Stevens (2007) for a more detailed discussion.

### 7.3.2 Naming scheme

SAS datasets with zero observations are attached to this document:[1]

- qwi/zz/qwi_zz_seinunit_d.sas7bdat
- qwi/zz/qwi_zz_seinunit_rh.sas7bdat
- qwi/zz/qwi_zz_seinunit_sa.sas7bdat
- qwi/zz/qwi_zz_seinunit_se.sas7bdat

`ZZ` stands for the state postal abbreviation, and `YYYY` for a calendar year. `_sa` identifies the sex-age tabulations, `_se` the sex-education tabulations, and `_rh` the race-ethnicity tabulations.

### 7.3.3 Data location

The files are stored in a main directory, with state-specific subdirectories:

```
qwi/ZZ/
```

[1] Also visible on the attachment tab - Adobe Reader may be required.

### 7.3.4 Main dataset: QWI_ZZ_SEINUNIT_AGG

The `QWI_ZZ_SEINUNIT_AGG` file (LEHD internal name: UFFb) is a file at the SEINUNIT level, providing detailed statistics for an establishment (SEIN + SEINUNIT) at every combination (AGG) of (SA) SEX x AGEGROUP, (SE) SEX x EDUCGROUP, (RH) RACE x ETHNICITY, or (D) no demographic groups. Age groups are defined using the WIA categorization. The different margins are represented as variable arrays in the UFFb.

Due to the very large number of variables, we only list some of the variables for the SA file. Zero-obs datasets are attached to this PDF for all files.

The generic variable name is constructed as `<STATISTIC>_<MARGIN1><MARGIN2>` where <STATISTIC> is one of the statistics described in Abowd et al. (2009), listed in Table 7.4 on page 7-9, and the values for the margins are taken from two "legal" combinations of codes from Table 7.5 on page 7-11. Thus, `A_1A02` are accessions $A$ for $sex = 1$ (men) of $agegroup = A02$ (ages 19-21) (on file `qwi_zz_seinunit_sa`), whereas `S_A4A1` are separations $S$ for Asians ($race = A4$) of non-Hispanic ethnicity ($ethnicity = A1$) (on file `qwi_zz_seinunit_rh`). Note that `<MARGIN1><MARGIN2>`=$GT0$ for all variables on `qwi_zz_seinunit_d`.

**Record identifier** YEAR QUARTER SEIN SEINUNIT
**Sort order** YEAR QUARTER SEIN SEINUNIT
**Entity** Establishment
**Unique Entity Key** SEIN SEINUNIT

Table 7.2: QWI_ZZ_SEINUNIT_SA: Extract of Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| . . . | | | |
| **A_0A00** | Accessions for All Sexes and age All Ages (14-99) | Num | 4 |
| **A_0A01** | Accessions for All Sexes and age 14-18 | Num | 4 |
| **A_0A02** | Accessions for All Sexes and age 19-21 | Num | 4 |
| **A_0A03** | Accessions for All Sexes and age 22-24 | Num | 4 |
| . . . | | | |
| **A_1A00** | Accessions for Male and age All Ages (14-99) | Num | 4 |
| **A_1A01** | Accessions for Male and age 14-18 | Num | 4 |
| **A_1A02** | Accessions for Male and age 19-21 | Num | 4 |
| . . . | | | |
| **A_2A00** | Accessions for Female and age All Ages (14-99) | Num | 4 |
| **A_2A01** | Accessions for Female and age 14-18 | Num | 4 |
| . . . | | | |
| **FS_1A03** | Flow out of full-quarter employment for Male and age 22-24 | Num | 4 |
| . . . | | | |
| . . . | | | |
| **FS_0A03** | Flow out of full-quarter employment for All Sexes and age 22-24 | Num | 4 |
| **FS_0A04** | Flow out of full-quarter employment for All Sexes and age 25-34 | Num | 4 |
| . . . | | | |

(cont.)

Table 7.2 (cont.): QWI_ZZ_SEINUNIT_SA: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **fdot_2A08** | Alternate definition of F that does not reflect flow suppression for Female and age 65-99 | Num | 4 |
| **leg_county** | Cleaned GEO FIPS County CCC | Char | 5 |
| **leg_subctygeo** | Sub-county geocode | Char | 10 |
| **leg_wib** | WIB code, wwwwww | Char | 6 |
| **naics2012fnl** | Final 2012 NAICS Code NNNNNN | Char | 6 |
| **quarter** | Quarter QQ | Num | 3 |
| **qwi_final_weight** | qwi_wcf*qwi_unit_weight | Num | 8 |
| **qwi_unit_weight** | Weight such that weighted sum of B_UI = sum(month1_BLS) | Num | 8 |
| **qwi_wcf** | QWI weight correction factor | Num | 8 |
| **sein** | State Employer ID Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **sic1987fnl** | Cleaned SIC Code IIII | Char | 6 |
| **unit_detail_flag** | =0 from ECF_SEIN, =1 if from ECF_SEINUNIT, =z not found | Char | 1 |
| **unit_quarters_off** | Number of quarters away from data that establishment was found | Num | 3 |
| **year** | Year YYYY | Num | 3 |

### 7.3.5 Main dataset: QWI_ZZ_SEINUNIT_D

The QWI_ZZ_SEINUNIT_D file (LEHD internal name: UFFb) is a file at the SEINUNIT level, providing detailed statistics for an establishment (SEIN + SEINUNIT), without any demographic breakout. This corresponds to the margins of the (AGG) files described above.

The generic variable name is constructed as <STATISTIC>_<GT0> where <STATISTIC> is one of the statistics listed in Table 7.4. Thus, A_GT0 are accessions $A$ for all ages, men and women, all race groups, all ethnicities, and all levels of education. In other words, it is simply the establishment-level count of accessions.

**Record identifier** YEAR QUARTER SEIN SEINUNIT
**Sort order** YEAR QUARTER SEIN SEINUNIT
**Entity** Establishment
**Unique Entity Key** SEIN SEINUNIT

Table 7.3: QWI_ZZ_SEINUNIT_D: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **AR_GT0** | Average accession rate | Num | 4 |
| **A_GT0** | Accessions | Num | 4 |
| **B_GT0** | Beginning-of-period employment | Num | 4 |
| **CA_GT0** | Flow into consecutive quarter employment | Num | 4 |
| **CH_GT0** | New Hires into Continuous Quarter Employment | Num | 4 |
| **CR_GT0** | Recalls into Continuous Quarter Employment | Num | 4 |
| **CS_GT0** | Flow out of consecutive quarter employment | Num | 4 |
| **E_GT0** | End-of-period employment | Num | 4 |
| **Ebar_GT0** | Average employment | Num | 4 |
| **FA_GT0** | Flow into full-quarter employment | Num | 4 |
| **FH_GT0** | Full-quarter new hires | Num | 4 |
| **FJCR_GT0** | Average full-quarter job creation rate | Num | 4 |
| **FJC_GT0** | Full-quarter job creation | Num | 4 |
| **FJD_GT0** | Full-quarter job destruction | Num | 4 |
| **FJF_GT0** | Net change in full-quarter employment | Num | 4 |
| **FS_GT0** | Flow out of full-quarter employment | Num | 4 |
| **FSnx_GT0** | Flow out of full-quarter employment - next quarter | Num | 4 |
| **F_GT0** | Full-quarter employment | Num | 4 |
| **Fbar_GT0** | Average full-quarter employment | Num | 4 |
| **Fpv_GT0** | Full-quarter employment - previous quarter | Num | 4 |
| **H_GT0** | New hires | Num | 4 |
| **JC_GT0** | Job creation | Num | 4 |
| **JD_GT0** | Job destruction | Num | 4 |

(cont.)

**Table 7.3 (cont.): QWI_ZZ_SEINUNIT_D: Variables and Attributes**

| Variable | Label | Type | Length |
|---|---|---|---|
| **JF_GT0** | Net job flows | Num | 4 |
| **M_GT0** | Employment any time during the period | Num | 4 |
| **R_GT0** | Recalls | Num | 4 |
| **SR_GT0** | Average separation rate | Num | 4 |
| **S_GT0** | Separations | Num | 4 |
| **W1_GT0** | Total payroll of all employees | Num | 4 |
| **ZW1_GT0** | Average monthly earnings of beginning-of-period employees | Num | 4 |
| **ZW3_GT0** | Average monthly earnings of full-quarter employees | Num | 4 |
| **ZWFA_GT0** | Average monthly earnings of transits to full-quarter status | Num | 4 |
| **ZWFH_GT0** | Average monthly earnings of new hires to full-quarter status | Num | 4 |
| **ZWFS_GT0** | Average monthly earnings of separations from full-quarter status | Num | 4 |
| **bdot_GT0** | Alternate definition of B that does not reflect flow suppression | Num | 4 |
| **edot_GT0** | Alternate definition of E that does not reflect flow suppression | Num | 4 |
| **es_owner_code** | Cleaned OWNER_CODE O | Char | 1 |
| **es_state** | ES202 FIPS State SS | Char | 2 |
| **fdot_GT0** | Alternate definition of F that does not reflect flow suppression | Num | 4 |
| **leg_county** | Cleaned GEO FIPS County CCC | Char | 5 |
| **leg_subctygeo** | Sub-county geocode | Char | 10 |
| **leg_wib** | WIB code, wwwwww | Char | 6 |
| **naics2012fnl** | Final 2012 NAICS Code NNNNNN | Char | 6 |
| **quarter** | Quarter QQ | Num | 3 |
| **qwi_final_weight** | qwi_wcf*qwi_unit_weight | Num | 8 |
| **qwi_unit_weight** | Weight such that weighted sum of B_UI = sum(month1_BLS) | Num | 8 |
| **qwi_wcf** | QWI weight correction factor | Num | 8 |
| **sein** | State Employer ID Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **sic1987fnl** | Cleaned SIC Code IIII | Char | 6 |
| **unit_detail_flag** | =0 from ECF_SEIN, =1 if from ECF_SEINUNIT, =z not found | Char | 1 |
| **unit_quarters_off** | Number of quarters away from data that establishment was found | Num | 3 |
| **year** | Year YYYY | Num | 3 |

### 7.3.6 Additional information

Table 7.4: Correspondence: public-use QWI variables and internal prefix

The following table shows the correspondence between public-use variables names on published QWI and internal prefixes, as well as the legal values for <STATISTIC> in the construction of variable names.

| Indicator variable | Alternate name | <STATISTIC> | Indicator name |
|---|---|---|---|
| Emp | B | B | Beginning-of-Quarter Employment |
| | | Bdot | Alternate definition of B that does not reflect flow suppression |
| EmpEnd | E | E | End-of-Quarter Employment |
| | | Edot | Alternate definition of E that does not reflect flow suppression |
| | | Ebar | Average employment $(B+E)/2$ |
| EmpS | F | F | Full-Quarter Employment (Stable) |
| | | Fdot | Alternate definition of F that does not reflect flow suppression |
| | | Fbar | Average full-quarter employment $(F+Fpv)/2$ |
| EmpSpv | Fpv | Fpv | Full-Quarter Employment in the Previous Quarter |
| EmpTotal | M | M | Employment - Reference Quarter |
| HirA | A | A | Hires (All Accessions) |
| HirN | H | H | New Hires |
| HirR | R | R | Recall Hires |
| Sep | S | S | Separations (All) |
| HirAEnd | CA | CA | End-of-Quarter Hires |
| HirAEndR | CAR | **n.a.** | End-of-Quarter Hiring Rate, compute as $CA/Ebar$ |
| SepBeg | CS | CS | Beginning-of-Quarter Separations |
| SepBegR | CSR | **n.a.** | Beginning-of-Quarter Separation Rate, compute as $CS/Ebar$ |
| HirAS | FA | FA | Hires (All Hires into Full-Quarter Employment) |
| HirNS | FH | FH* | New Hires (New Hires into Full-Quarter Employment) |
| SepS | FS | FS | Separations (Flows out of Full-Quarter Employment) |
| SepSnx | FSnx | FSnx | Separations in the Next Quarter (Flows out of Full-Quarter Employment) |
| TurnOvrS | FT | **n.a.** | Turnover (Stable), compute as $(FA+FSnx)/(2F)$ |
| FrmJbGn | JC | JC | Firm Job Gains (Job Creation) |
| FrmJbLs | JD | JD | Firm Job Loss (Job Destruction) |
| FrmJbC | JF | JF | Firm Job Change (Net Change) |
| HirAEndRepl | EI | **n.a.** | Replacement Hires, compute as $CA-JC$ |
| HirAEndReplr | EIR | **n.a.** | Replacement Hiring Rate, compute as $EI/Ebar$ |
| FrmJbGnS | FJC | FJC | Firm Job Gains (Stable) |

NOTE: A <STATISTIC> marked with "*" has been renamed relative to the internal naming on the UFF_B.

Table 7.4: Correspondence between public-use QWI variables and internal prefix (cont)

| Indicator<br>variable | Alternate<br>name | <STATISTIC> | Indicator<br>name |
|---|---|---|---|
| FrmJbLsS | FJD | FJD | Firm Job Loss (Stable) |
| FrmJbCS | FJF | FJF | Firm Job Change (Stable; Net Change) |
| EarnS | ZW3 | ZW3* | Average Monthly Earnings (Full-Quarter Employment) |
| EarnBeg | ZW1 | ZW1* | Average Monthly Earnings (Beginning-of-Quarter Employment) |
| EarnHirAS | ZWFA | ZWFA* | Average Monthly Earnings (All Hires into Full-Quarter Employment) |
| EarnHirNS | ZWFH | ZWFH* | Average Monthly Earnings (New Hires into Full-Quarter Employment) |
| EarnSepS | ZWFS | ZWFS* | Average Monthly Earnings (Flows out of Full-Quarter Employment) |
| Payroll | W1 | W1 | Total Quarterly Payroll |

NOTE: A <STATISTIC> marked with "*" has been renamed relative to the internal naming on the UFF_B.

Table 7.5: QWI coding

*QWI coding: Sex*

| sex | label |
|---:|---|
| 0 | Male and Female |
| 1 | Male |
| 2 | Female |

*QWI coding: Age groups*

| agegrp | label |
|---:|---|
| A00 | All Ages (14-99) |
| A01 | 14-18 |
| A02 | 19-21 |
| A03 | 22-24 |
| A04 | 25-34 |
| A05 | 35-44 |
| A06 | 45-54 |
| A07 | 55-64 |
| A08 | 65-99 |

*QWI coding: Education groups*

| education | label |
|---:|---|
| E0 | All Education Categories |
| E1 | Less than high school |
| E2 | High school or equivalent, no college |
| E3 | Some college or Associate degree |
| E4 | Bachelor’s degree or advanced degree |
| E5 | Educational attainment not available (workers aged 24 or younger) |

Table 7.5 (cont.)

*QWI coding: Ethnicity*

| ethnicity | label |
|---:|---|
| A0 | All (Any Ethnicity) |
| A1 | Not Hispanic or Latino |
| A2 | Hispanic or Latino |

*QWI coding: Race*

| race | label |
|---:|---|
| A0 | All Races |
| A1 | White Alone |
| A2 | Black or African American Alone |
| A3 | American Indian or Alaska Native Alone |
| A4 | Asian Alone |
| A5 | Native Hawaiian or Other Pacific Islander Alone |
| A6 | Some Other Race Alone (Not Used) |
| A7 | Two or More Race Groups |

Source: LEHD Schema V4.0.1 at lehd.ces.census.gov/data/schema/V4.0.1

### 7.3.7 Summary information on datasets

## 7.4 NOTES

# Chapter 8. Quarterly Workforce Indicators - Public-use files (QWIPU)

## 8.1 OVERVIEW

The public-use Quarterly Workforce Indicators (QWIPU) provide local labor market statistics by industry, worker demographics, employer age and size.Unlike statistics tabulated from firm or person-level data, the QWI source data are unique job-level data that link workers to their employers. Because of this link, labor market data in the QWI is available by worker age, sex, educational attainment, and race/ethnicity. This allows for analysis by demographics of a particular local labor market or industry - for instance, identifying industries with aging workforces at the county level. Indicators include employment levels and changes, as well as worker flows - hires, separations, and turnover - and job flows - job creation, destruction, and net employment growth. More detailed information is available at Quarterly Workforce Indicators 101, available online at http://lehd.ces.census.gov/doc/QWI_101.pdf.

## 8.2 DATA AVAILABILITY

Time is reported on the QWI by specifying a year and calendar quarter (Jan-Mar, Apr-Jun, Jul-Sep, Oct-Nov). The extent of the time series available will vary by state, based on the availability of historical data when joining the partnership. The earliest state time series begin in 1990.

### 8.2.1 QWI Data Releases

The QWI are produced on a quarterly schedule. In the event that data submission or data quality issues are encountered, QWI production for a state may be skipped for one or several quarters, until the issue can be resolved.

### 8.2.2 Updates and revisions

The complete QWI time series is recalculated with every release, so numbers may change in any quarter. These changes are due to a number of factors, including:

- Updates to input files (primarily UI and QCEW)
  - States typically make a second submission of the previous quarter's data in every quarter, to improve completeness of data reporting. Historical files may be resubmitted to improve data quality.
  - Other input data sources are also periodically updated.
- Modifications to algorithms to develop estimates
  - The data quality of the QWI is continuously reviewed, and the algorithms are periodically modified to improve the results. These modifications may affect measures throughout the time series.

- Stochastic changes to imputations used to complete missing information
  - Random draws are used to generate data that are missing. These draws may change between production runs, though longitudinal consistency is generally maintained within a data release.

For this reason, analyses using the public-use QWIPU should always reference the correct release of the data. The `version.txt` file contains the metadata of each state's release, and identifies the release version of the data, both on the internet as well as on the FSRDC. Each state will have their own `version.txt` file.

## 8.3 DATA AVAILABILITY IN THE FSRDC

Access to the QWIPU requires no additional permissions (public-use data). However, access to confidential LEHD microdata precludes access to public-use QWI, and vice-versa. Researchers must make a choice for the duration of their project.

- Researchers who only require local area or industry controls are well-served with the QWIPU.
- Researchers who require firm-level data may be well-served with the (confidential) establishment-level QWI (see Chapter 7).
- Researchers who require cross-firm linkages, job- or person-level data should request confidential LEHD microdata.

QWIPU are provided upon request, and are not provisioned automatically to the FSRDC. Please contact the FSRDC administrator on how to request the data. The data are provided as compressed CSV files, and need to be read-in by individual researchers. SAS readin programs are made available.

# Chapter 9. Successor-Predecessor file (SPF)

## 9.1 OVERVIEW

The Successor-Predecessor File (SPF) is a suite of files providing intertemporal flow-based links based on wage records and administrative links. The file is not fully documented, researchers are advised to use the file with caution.

## 9.2 DATA CITATION

U.S. Census Bureau. 2016. *Successor-Predecessor Files (SPF) in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

## 9.3 DETAILED DESCRIPTION

### 9.3.1 Definition of Successor-Predecessor

The successor-predecessor sequence creates two files, the SPF (Successor-Predecessor File) which has a record for every link (whether that link is identified by employee flows from the UI wage records or from the successor-predecessor data on the ES-202) between SEINs, and the WSLF (Within-SEIN Links File) which has a record for every successor-predecessor link reported on the ES-202 between SEINUNITs within the same SEIN.

### 9.3.2 Processing description

First, we read the PIK-SEIN work history information from the EHF into simple character strings of 1's or 0's referring to whether or not the PIK has positive earnings at the SEIN in the quarter corresponding to the position in the character string. We then match up each end of job string experienced by a PIK with the beginning of job strings for that PIK at another SEIN which start in the same or subsequent quarter that the first job ends. We then sum up the number of such flows between each SEIN pair in a given quarter. If the number of transitioning employees and the SEINs involved satisfy certain criteria, then a link is recorded for that SEIN pair in that quarter. We then read in the successor-predecessor information from the ES-202 and divide the data into a within-SEIN links file and an across-SEIN links file. The across-SEIN links file is aggregated to the SEIN-level for comparability to the links formed with the UI wage records. Finally, the UI wage record links and the SEIN-level, ES-202 links are merged into one file.

### 9.3.3 Changes in this Snapshot

None.

## 9.4 DATA SET DESCRIPTIONS

### 9.4.1 Naming scheme

All files start with `spf`. The main SPF file has no suffix, other files have a suffix. SAS datasets with zero observations are attached to this document:[1]

- spf/zz/spf_zz.sas7bdat
- spf/zz/spf_zz_wslf.sas7bdat

`ZZ` stands for the state postal abbreviation.

### 9.4.2 Data location

The files are stored in state-specific subdirectories of the main SPF directory:

```
spf/ZZ/
```

No files in the `spf` process contain Title 26 data.

[1] Also visible on the attachment tab - Adobe Reader may be required.

### 9.4.3 UI-based Output Files

#### 9.4.3.1 SPF

The main SPF stores links between SEINs within a state (no cross-state links). Key variables are `link_ui` and `link_es`.

**Record identifier** SEIN-SEIN_SUCC
**Sort order** SEIN-SEIN_SUCC
**Entity** Link between firms
**Unique Entity Key** SEIN

Table 9.1: SPF_ZZ: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **active_beg_qtr_a** | First quarter predecessor is active on UI | Num | 3 |
| **active_beg_qtr_b** | First quarter successor is active on UI | Num | 3 |
| **active_end_qtr_a** | Last quarter predecessor is active on UI | Num | 3 |
| **active_end_qtr_b** | Last quarter successor is active on UI | Num | 3 |
| **bpemp_master** | Pred UI B Employment, max of last 3 quarters | Num | 8 |
| **emp_es** | Pred ES202 Month 1 Employment, max of last three quarters | Num | 8 |
| **es_qtrs_off** | Number of quarters removed ES202 event date is from UI flow | Num | 3 |
| **link_es** | Type of ES202 based link | Num | 3 |
| **link_ui** | Type of link for predecessor firm | Num | 3 |
| **match_period** | Percent of transitions where separation precedes quarter of accession | Num | 8 |
| **num_left** | Number of jobs transitioning between firms | Num | 8 |
| **qtime** | Quarter of separation, 1985Q1=1 | Num | 8 |
| **ratio** | Percent of jobs at predecessor transitioning to successor (estimated) | Num | 8 |
| **sein** | SEIN - predecessor | Char | 12 |
| **sein_succ** | SEIN - successor | Char | 12 |
| **source** | Data source of link between firms | Char | 4 |
| **succ_link_ui** | Type of link for successor firm | Num | 3 |
| **succ_ratio** | Percent of jobs at successor transitioning from predecessor (estimated) | Num | 8 |

### Values taken by UI link variables

```
/*   +--< LEHD-QWI spf 3.1.25 2005-04-21 schwa305            >--+   */
/*   +--< Location: /programs/production/dev1/current/spf    >--+   */
/*   +--< File: library/formats/links_ui.sas                 >--+   */
/* Time-stamp: <04/10/20 17:58:32 vilhuber> */
/*BEGINCCC
    Format created to tabulate the variable LINK\_UI.

CCCEND*/

proc format;
    value linkui
1="Pred exits & 80% Pred Employment moves to Succ        "
2="Pred exits & <80% Pred Employment moves to Succ       "
3=""
4="Pred does not exit & 80% Pred Employment moves to Succ"
5="Pred does not exit & <80% Pred Employment moves to Scc"
6="" ;
run;
```

### Values taken by Successor UI link variables

```
/*   +--< LEHD-QWI spf 3.1.25 2005-04-21 schwa305            >--+   */
/*   +--< Location: /programs/production/dev1/current/spf    >--+   */
/*   +--< File: library/formats/succ_link_ui.sas             >--+   */
/* Time-stamp: <04/10/20 18:04:26 vilhuber> */
/*BEGINCCC
    Format created to tabulate the variable SUCC\_LINK\_UI.

CCCEND*/

proc format;
    value slinkui
1="Succ enters & 80% Succ Employment comes from Pred        "
2="Succ enters & <80% Succ Employment comes from Pred       "
3=""
4="Succ does not enter & 80% Succ Employment comes from Pred"
5="Succ does not enter & <80% Succ Employment comes from Pred"
6="" ;
run;
```

#### 9.4.3.2 SPF-WSLF

**Record identifier** PIK-SEIN-SEINUNIT
**Sort order** PIK-SEIN-SEINUNIT
**Entity** Job
**Unique Entity Key** PIK-SEIN-SEINUNIT

Table 9.2: SPF_ZZ_WSLF: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **pred_sein** | Predecessor SEIN | Char | 12 |
| **pred_seinunit** | Predecessor SEINUNIT | Char | 5 |
| **quarter** | Quarter QQ | Num | 4 |
| **sein** | State Employer ID Number | Char | 12 |
| **seinunit** | State UI Reporting Unit Number | Char | 5 |
| **succ_sein** | Successor SEIN | Char | 12 |
| **succ_seinunit** | Successor SEINUNIT | Char | 5 |
| **year** | Year YYYY | Num | 5 |

### 9.4.4 Summary information on datasets

## 9.5 NOTES

# Chapter 10. Unit-to-Worker Impute - Job location impute (U2W)

## 10.1 OVERVIEW

The UI records underlying the LEHD Infrastructure files provide neither establishment identifiers (except for Minnesota), nor industry or geographic detail of the establishment, only a firm identifier. Between 60 and 70 percent of state-level employment is in single-unit employers (employers with only one establishment), for which a link through the firm identifier is sufficent to provide such detail. For the remaining 30 to 40 percent of employment, such links have to be imputed. The Unit-to-Worker Impute (U2W) file contains ten imputed establishments for each employee of a multi-unit employer. The file can be linked to other Census Bureau datasets through the PIK and the LEHD SEIN-SEINUNIT.

### 10.1.1 Changes in this Snapshot

None.

## 10.2 DATA CITATION

U.S. Census Bureau. 2016. *Unit-to-Worker Impute (U2W) files in LEHD Infrastructure, S2014 Version.* [Computer file]. Washington,DC: U.S. Census Bureau, Center for Economic Studies, Research Data Centers [distributor].

## 10.3 DETAILED DESCRIPTION

The information in this section draws heavily on Abowd et al. (2009) and Stephens (2007).

A primary objective of the QWI is to provide employment, job and worker flows, and wage measures at a very detailed levels of geography (place-of-work) and industry. The structure of the administrative data received by LEHD from state partners, however, poses a challenge to achieving this goal. QWI measures are primarily based on the processing of UI wage records which report, with the exception of Minnesota, only the legal employer (SEIN) of the workers. The ES-202 micro-data, however, are comprised of establishment-level records which provide the geographic and industry detail needed to produce the QWI. For employers operating only one establishment within a state, the assignment of establishment-level characteristics to UI wage records is straightforward because there is no distinction between the employer and the establishment. However, approximately 30 to 40 percent of state-level employment is concentrated in employers that operate more than one establishment in that state. For these multi-unit employers, the SEIN on workers' wage records identifies the legal employer in the ES-202 data, but not the employing establishment (place-of-work). Thus, establishment level characteristics–geography and industry, in particular–are missing data for these multi-unit job histories.

In order to impute establishment-level characteristics to job histories of multi-unit employers, a non-ignorable missing data model with multiple imputation was developed. The model imputes establishment-of-employment using two key characteristics available in the LEHD Infrastructure Files: 1) distance between place-of-work and place-of-residence and 2) the distribution of employment across establishments of multi-unit employers. The distance to work model is estimated using data from Minnesota, where both the SEIN and SEINUNIT identifiers appear on a UI wage record. Then, the posterior distribution of the parameters from this estimation, combined with the actual SEIN and SEINUNIT employment histories from the ES-202 data, are used for multiple imputation of the SEINUNIT associated with for workers in a given SEIN in the data from states other than Minnesota.[1] Emerging from this process is an output file, called the Unit-to-Worker (U2W) file, containing ten imputed establishments for each worker of a multi-unit employer. These implicates are then used in the downstream processing of the QWI.

The U2W process relies on information from each of the four Infrastructure Files–ECF, GAL, EHF, and ICF–as well as the auxiliary SPF file. Within the ECF, the universe of multi-unit employers is identified. For these employers, the ECF also provides establishment-level employment, date-of-birth, and geocodes (which are acquired from the GAL). The SPF contains information on predecessor relationships which may lead to the revision of date-of-birth implied by the ECF. Finally, job histories in the EHF in conjunction with place-of-residence information stored in the ICF provide the necessary worker information needed to estimate and apply the imputation model.

## 10.3.1 A probability model for employment location

### 10.3.1.1 Definitions

Let $i = 1, ..., I$ index workers, $j = 1, ..., J$ index employers (SEINs), and $t = 1, ..., T$ index time (quarters). Let $R_{jt}$ denote the number of active establishments at employer $j$ in quarter $t$, let $\mathfrak{R} = \max_{j,t} R_{jt}$, and $r = 1, ..., \mathfrak{R}$ index establishments. Note that the index $r$ is nested within $j$. Let $N_{jrt}$ denote the quarter $t$ employment of establishment $r$ in employer $j$. Finally, if worker $i$ was employed at employer $j$ in $t$, denote by $y_{ijt}$ the establishment at which the worker was employed.

Let $\mathcal{J}_t$ denote the set of employers active in quarter $t$, let $\mathcal{I}_{jt}$ denote the set of individuals employed at employer $j$ in quarter $t$, let $\mathcal{R}_{jt}$ denote the set of active ($N_{jrt} > 0$) establishments at employer $j$ in $t$, and let $\mathcal{R}^i_{jt} \subset \mathcal{R}_{jt}$ denote the set of active establishments that are feasible for worker $i$. Feasibility is defined as follows. An establishment $r \in \mathcal{R}^i_{jt}$ if $N_{jrs} > 0$ for every quarter $s$ that $i$ was employed at $j$.

### 10.3.1.2 The probability model

Let $p_{ijrt} = \Pr(y_{ijt} = r)$. At the core of the model is the probability statement:

$$p_{ijrt} = \frac{e^{\alpha_{jrt} + x'_{ijrt}\beta}}{\sum_{s \in \mathcal{R}^i_{jt}} e^{\alpha_{jst} + x'_{ijst}\beta}} \tag{10.1}$$

where $\alpha_{jrt}$ is a establishment- and quarter-specific effect, $x_{ijrt}$ is a time-varying vector of characteristics of the worker and establishment, and $\beta$ measures the effect of characteristics on the probability of being employed at a particular establishment. In the current implementation, $x_{ijrt}$ is a linear spline in the (great-circle) distance between worker $i$'s residence and the physical location of establishment $r$. The spline has knots at 25, 50, and 100 miles.

[1] The actual SEINUNIT coded on the UI wage records is used for Minnesota, and would be used for any other state that provided such data. Note that there are occasional, and rare, discrepancies between the unit structure on the Minnesota wage records and the unit structure on the Minnesota ES-202 data for the same quarter. These discrepancies are resolved during the initial processing of the Minnesota data in its state-specific readin procedures.

Using (10.1), the following likelihood is defined

$$p\left(y|\alpha,\beta,x\right) = \prod_{t=1}^{T} \prod_{j\in\mathcal{J}_t} \prod_{i\in\mathcal{I}_{jt}} \prod_{r\in\mathcal{R}^i_{jt}} \left(p_{ijrt}\right)^{d_{ijrt}} \tag{10.2}$$

where

$$d_{ijrt} = \begin{cases} 1 & \text{if } y_{ijt} = r \\ 0 & \text{otherwise} \end{cases} \tag{10.3}$$

and where $y$ is the appropriately-dimensioned vector of the outcome variables $y_{ijt}$, $\alpha$ is the appropriately-dimensioned vector of the $\alpha_{jrt}$, and $x$ is the appropriately-dimensioned matrix of characteristics $x_{ijrt}$. For $\alpha_{jrt}$, a hierarchical Bayesian model based on employment counts $N_{jrt}$ is specified.

The object of interest is the joint posterior distribution of $\alpha$ and $\beta$. A uniform prior on $\beta$, $p\left(\beta\right) \propto 1$ is assumed. The characterization of $p\left(\alpha,\beta|x,y,N\right)$ is based on the factorization

$$\begin{aligned} p\left(\alpha,\beta|x,y,N\right) &= p\left(\alpha|N\right)p\left(\beta|\alpha,x,y\right) \\ &\propto p\left(\alpha|N\right)p\left(\beta\right)p\left(y|\alpha,\beta,x\right) \\ &\propto p\left(\alpha|N\right)p\left(y|\alpha,\beta,x\right). \end{aligned} \tag{10.4}$$

Thus, the joint posterior (10.4) is completely characterized by the posterior of $\alpha$ and the likelihood of $y$ in (10.2). Note (10.2) and (10.4) assume that the employment counts $N$ affect employment location $y$ only through the parameters $\alpha$.

#### 10.3.1.3 Estimation

The joint posterior $p\left(\alpha,\beta|x,y,N\right)$ is approximated at the posterior mode. In particular, we estimate the posterior mode of $p\left(\beta|\alpha,x,y\right)$ evaluated at the posterior mode of $\alpha$. From these we compute the posterior modal values of the $\alpha_{jrt}$, then, maximize the log posterior density

$$\log p\left(\beta|\alpha,x,y\right) \propto \sum_{t=1}^{T} \sum_{j\in\mathcal{J}_t} \sum_{i\in\mathcal{I}_{jt}} \sum_{r\in\mathcal{R}^i_{jt}} d_{ijrt}\left(\alpha_{jrt} + x'_{ijrt}\beta - \log\left(\sum_{s\in\mathcal{R}^i_{jt}} e^{\alpha_{jst}+x'_{ijst}\beta}\right)\right) \tag{10.5}$$

which is evaluated at the posterior modal values of the $\alpha_{jrt}$, using a modified Newton-Raphson method. The mode-finding exercise is based on the gradient and Hessian of (10.5). In practice, (10.5) is estimated for three employer employment size classes: 1-100 employees, 101-500 employees, and greater than 500 employees, using data for Minnesota.

### 10.3.2 Imputing place of work

After estimating the probability model using Minnesota data, the posterior distribution of the estimated $\beta$ parameters is combined with the entity specific posterior distribution of the $\alpha$ parameters in the imputation process for other states. A brief outline of the imputation method, as it relates to the probability model previously discussed, is provided in this section. Emphasis is placed on not only the imputation process itself, but also the preparation of input data.

#### 10.3.2.1 Sketch of the imputation method

Ignoring temporal considerations, 10 implicates are generated as follows. First, using the posterior mean and variance of $\beta$ estimated from the Minnesota data, we take 10 draws of $\beta$ from the normal approximation (at the mode) to $p\left(\beta|\alpha,x,y\right)$. Next, using ES-202 employment counts for the establishments, we compute 10 values of $\alpha_{jt}$ based on the hierarchical model for these parameters. Note that these are draws from the exact posterior distribution of the $\alpha_{jrt}$. The drawn values of $\alpha$ and $\beta$ are used to draw 10 imputed values of place of work from the asymptotic approximation to the posterior predictive distribution

$$p\left(\tilde{y}|x,y\right)=\int\int p\left(\tilde{y}|\alpha,\beta,x,y\right)p\left(\alpha|N\right)p\left(\beta|\alpha,x,y\right)d\alpha d\beta. \tag{10.6}$$

#### 10.3.2.2 Implementation

**Establishment data** Using state-level micro-data, the set of employers (SEINs) that ever operate more that one establishment in a given quarter is identified; these SEINs represent the set of ever-multi-unit employers defined above as the set $\mathcal{J}_t$. For each of these employers, its establishment-level records are identified. For each establishment, latitude and longitude coordinates, parent employer (SEIN) employment, and ES-202 month-one employment[2] for the entire history of the establishment are retained. Those establishments with positive month-one employment in a given quarter characterize $\mathcal{R}_{jt}$, the set of all active establishments. An establishment birth date is identified and, in most cases, is the first quarter in the ES-202 time series in which the establishment has positive month-one employment. For some employers, predecessor relationships are identified in the SPF; in those instances, the establishment date-of-birth is adjusted to coincided with that of the predecessor's.

**Worker data** The EHF provides the earnings histories for employees of the ever-multi-unit employers. For each in-scope job (a worker-employer pair), one observation is generated for the *end* of each job spell, where a job spell is defined as a continuum of quarters of positive earnings for worker at a particular employer during which there are no more than 3 consecutive periods of non-positive earnings.[3] The start date of the job history is identified as the first quarter of positive earnings; the end date is the last date of positive earnings.[4] These job spells characterize the set $\mathcal{I}_{jt}$

**Candidates** Once the universe of establishments and workers is identified, data are combined and *a priori* restrictions and feasibility assumptions are imposed. For each quarter of the date series, the history of every job spell that *ends in that quarter* is compared to the history of *every* active (in terms of ES-202 first month employment) establishment of the employing employer (SEIN). The start date of the job spell is compared to the birth date of each establishment. Establishments that were born after the start of a job spell are immediately discarded from the set of candidate establishments. The remaining establishments constitute the set $\mathcal{R}_{jt}^{i}\subset\mathcal{R}_{jt}$ for a job spell (worker) at a given employer.[5]

Given the structure of the pairing of job spells with candidate establishments, it is clear that within job spell changes of establishment are ruled-out. An establishment is imputed once for each job spell,[6] thereby creating no spurious labor market transitions.

**Imputation and output data** Once the input data are organized, a set of 10 imputed establishment identifiers are generated for each job spell ending in every quarter for which both ES-202 and UI wage records exist. For each quarter, implicate, and size class, $s=1,2,3$, the parameters on the linear spline in distance between place-of-work

[2] In rare instances where no ES-202 employment is available, an alternative employment measure based on UI wage record counts may be used.

[3] A new hire is defined in the QWI as a worker who accedes to a firm in the current period but was not employed by the same firm in any of the 4 previous periods. A new job spell is created if, for example, a worker leaves a firm for more than 4 quarters and is subsequently re-employed by the same firm.

[4] By definition, an end-date for a job spell is not assigned in cases where a quarter of positive earnings at a firm is succeeded by 4 or fewer quarters of non-employment and subsequent re-employment by the same firm.

[5] The sample of UI wage and QCEW data chosen for processing of the QWI is such that the start and end dates are the same. Birth and death dates of establishments are, more precisely, the dates associated with the beginning and ending of employment activity observed in the data. The same is true for the dates assigned to the job spells.

[6] More specifically, an establishment is imputed to a job spell only once within each implicate.

and place-of-residence $\hat{\beta}^s$ are sampled from the normal approximation of the posterior predictive distribution of $\beta^s$ conditional on Minnesota ($MN$)

$$p(\beta^s|\alpha_{MN}, x_{MN}, y_{MN}) \tag{10.7}$$

The draws from this distribution vary across implicates, but not across time, employers, and individuals. Next, for each employer $j$ at time $t$, a set of $\hat{\alpha}_{jrt}$ are drawn from

$$p\left(\alpha_{ST}|N_{ST}\right) \tag{10.8}$$

which are based on the ES-202 month-one employment totals ($N_{jrt}$) for all candidate establishments $r_{jt} \subset \mathcal{R}_{jt}$at employer $j$ within the state ($ST$) being processed. The initial draws of $\hat{\alpha}_{jrt}$ from this distribution vary across time and employers but not across job spells. Combining (10.7) and (10.8) yields

$$\begin{aligned} & p\left(\alpha_{ST}|N_{ST}\right) p(\beta^s|\alpha_{MN}, x_{MN}, y_{MN}) \\ \approx \; & p\left(\alpha_{ST}|N_{ST}\right) p(\beta^s|\alpha_{ST}, x_{ST}, y_{ST}) \\ = \; & p\left(\alpha_{ST}, \beta_{ST}|x_{ST}, y_{ST}, N_{ST}\right), \end{aligned} \tag{10.9}$$

an approximation of the joint posterior distribution of $\alpha$ and $\beta^s$ (10.4) conditional on data from the state being processed.

The draws $\hat{\beta}^s$ and $\hat{\alpha}_{jrt}$ in conjunction with the establishment, employer, and job spell data are used to construct the $p_{ijrt}$ in (10.1) for all candidate establishments $r \in \mathcal{R}^i_{jt}$. For each job spell and candidate establishment combination, the $\hat{\beta}^s$ are applied to the calculated distance between place-of-residence (of the worker holding the job spell) and the location of the establishment, where the choice of $\hat{\beta}^s$depends on the size class of the establishment's parent employer. For each combination an $\hat{\alpha}_{jrt}$ is drawn which is based primarily on the size (in terms of employment) of the establishment relative to other active establishments at the parent employer. In conjunction, these determine the conditional probability $p_{ijrt}$ of a candidate establishment's assignment to a given job spell. Finally, from this distribution of probabilities is drawn an establishment of employment.

The imputation process yields a data file containing a set of 10 imputed establishment identifiers for each job spell. In a very small set of cases, the model fails to impute an establishment to a job spell. This is often due to unanticipated idiosyncrasies in the underlying administrative data. Furthermore, across states, the proportion of these failures relative to successful imputation is well under 0.5%. For these job spells, a dummy establishment identifier is assigned and in downstream processing, the employment-weighted modal employer-level characteristics are used.

## 10.4 DATA SET DESCRIPTIONS

### 10.4.1 Naming scheme

The U2W contains a single file per state: SAS datasets with zero observations are attached to this document:[7]

- u2w/zz/u2w_zz.sas7bdat

ZZ stands for the state postal abbreviation.

### 10.4.2 Data location

The files are stored in a main directory, with state-specific subdirectories:

```
u2w/ZZ/
```

[7] Also visible on the attachment tab - Adobe Reader may be required.

### 10.4.3 Main dataset: u2w_zz

This files contain the 10 imputed establishment identifiers are generated for each job spell.

**Record identifier** PIK SEIN NEW_HIST_FLAG
**Sort order** PIK SEIN NEW_HIST_FLAG
**Entity** Job spell
**Unique Entity Key** PIK SEIN

Table 10.1: U2W_ZZ: Variables and Attributes

| Variable | Label | Type | Length |
|---|---|---|---|
| **first_date** | Start of spell YYYY.F (e.g. 2000Q2=2000.25) | Num | 3 |
| **imputed_unit_1** | State UI Reporting Unit Number (Impute 1) | Char | 5 |
| **imputed_unit_2** | State UI Reporting Unit Number (Impute 2) | Char | 5 |
| **imputed_unit_3** | State UI Reporting Unit Number (Impute 3) | Char | 5 |
| **imputed_unit_4** | State UI Reporting Unit Number (Impute 4) | Char | 5 |
| **imputed_unit_5** | State UI Reporting Unit Number (Impute 5) | Char | 5 |
| **imputed_unit_6** | State UI Reporting Unit Number (Impute 6) | Char | 5 |
| **imputed_unit_7** | State UI Reporting Unit Number (Impute 7) | Char | 5 |
| **imputed_unit_8** | State UI Reporting Unit Number (Impute 8) | Char | 5 |
| **imputed_unit_9** | State UI Reporting Unit Number (Impute 9) | Char | 5 |
| **imputed_unit_10** | State UI Reporting Unit Number (Impute 10) | Char | 5 |
| **last_date** | End of spell YYYY.F (e.g. 2000Q4=2000.75) | Num | 3 |
| **new_hist_flag** | Spell number for same SEIN | Num | 3 |
| **pik** | Protected Identification Key | Char | 9 |
| **sein** | State Employer Identification Number | Char | 12 |

## 10.5 NOTES

# Appendix A. Acronyms used

**ACS-POW** American Community Survey Place of Work file

**AHS** American Housing Survey

**BED** Business Employment Dynamics

**BII** Business-identifying information

**BLS** Bureau of Labor Statistics

**BMF** Block Map File

**BR** Business Register, formerly known as the SSEL

**BRB** Business Register Bridge

**CBSA** Core-Based Statistical Area

**CEW** Covered Employment and Wages

**COLA** cost of living allowance

**CPR** Composite Person Record

**CPS** Current Population Survey

**DC** Decennial Census

**DHS** Department of Homeland Security

**DRB** Disclosure Review Board

**ECF** Employer Characteristics File

**ES-202** ES-202. An older name for the QCEW program

**EHF** Employment History Files

**EHRI** Enterprise Human Resources Integration

**EIN** (federal) Employer Identification Number

**FBI** Federal Bureau of Investigation

**FEMA** Federal Emergency Management Agency

**FIPS** Federal Information Processing Standards codes issued by National Institute of Standards and Technology (NIST)

**FTI** Federal Tax Information, typically covered under Title 26, U.S.C.

**FSRDC** Federal Statistical Research Data Center

**GAL** Geocoded Address List

**GSA** General Services Administration

**GRF-C** Geographic Reference File-Codes

**GRF-C** Geographic Reference File - Codes

**HHS** Department of Health and Human Services

**ICF** Individual Characteristics File

**IRS** Internal Revenue Service

**IRS** Internal Revenue Service

**JHF** Job History File

**LBD** Longitudinal Business Database

**LED** Local Employment Dynamics

**LEHD** Longitudinal Employer-Household Dynamics

**LMI** Labor Market Information

**LODES** LEHD Origin-Destination Employment Statistics

**MAF** Master Address File

**MN** Minnesota

**MOU** Memorandum of Understanding

**MSA** Metropolitan Statistical Area

**NAICS** North American Industry Coding System

**NIST** National Institute of Standards and Technology

**OMB** Office of Management and Budget

**OPM** Office of Personnel Management

**OTM** OnTheMap

**PHF** Person History File

**PIK** Protected Identity Key

**POI** Point of informationfile, one of the OPM data files

**QCEW** Quarterly Census of Employment and Wages, managed by the Bureau of Labor Statistics (BLS)

**QWI** Quarterly Workforce Indicators

**QWIPU** Public-Use QWI

**RDC** Research Data Center

**SCT** Standard Code Table, one of the OPM data files

**SEIN** State employer identification number. It is constructed from the state Federal Information Processing Standards (FIPS) code and the UI account number. The BLS refers to the UI account number in combination with the reporting unit number as SESA-ID

**SEINUNIT** SEIN reporting unit

**SESA** State Employment Security Agency

**SIC** Standard Industry Classification

**SPF** Successor-Predecessor File
**SSEL** Standard Statistical Establishment List
**SSN** Social Security Number
**U2W** Unit-to-Worker Impute
**UI** unemployment insurance
**WIB** Workforce Investment Board

# Bibliography

Abowd, John M., Bryce E. Stephens, and Lars Vilhuber. 2006. *Confidentiality Protection in the Census Bureau's Quarterly Workforce Indicators.* Technical Paper TP-2006-02. LEHD, U.S. Census Bureau. http://econpapers.repec.org/paper/centpaper/2006-02.htm.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon D. Woodcock. 2006. *The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators.* Technical report TP-2006-01. U.S. Census Bureau, LEHD and Cornell University. http://econpapers.repec.org/paper/centpaper/2006-01.htm.

———. 2009. "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators." In *Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts,* edited by Timothy Dunne, J. Brad Jensen, and Mark J. Roberts. University of Chicago Press. http://www.nber.org/chapters/c0485.pdf.

American Economic Association. 2014. "The American Economic Review: Data Availability Policy." Accessed April 16, 2014. http://www.aeaweb.org/aer/data.php.

Benedetto, Gary, John Haltiwanger, Julia Lane, and Kevin McKinney. 2007. "Using Worker Flows in the Analysis of the Firm." *Journal of Business and Economic Statistics* 25, no. 3 (July): 299–313.

DeSalvo, Bethany, Frank F. Limehouse, and Shawn D. Klimek. 2016. *Documenting the Business Register and Related Economic Business Data.* Working Papers 16-17. Center for Economic Studies, U.S. Census Bureau, March. https://ideas.repec.org/p/cen/wpaper/16-17.html.

Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. 2017. "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics." In *Proceedings of the 2017 International Conference on Management of Data,* vol. forthcoming. SIGMOD '17. ACM. doi:10.1145/3035918.3035940. http://dx.doi.org/10.1145/3035918.3035940.

Journal of Labor Economics. 2009. "Data Policy." Accessed April 16, 2014. http://www.press.uchicago.edu/journals/jole/datapolicy.html?journal=jole.

Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. "Privacy: Theory meets practice on the map." *International Conference on Data Engineering (ICDE)* (Washington, DC, USA): 277–286. doi:10.1109/ICDE.2008.4497436. http://dx.doi.org/10.1109/ICDE.2008.4497436.

McKinney, Kevin, and Lars Vilhuber. 2011a. *LEHD Data Documentation LEHD-OVERVIEW-S2008-rev1.* Working Papers 11-43. Center for Economic Studies, U.S. Census Bureau, December. http://ideas.repec.org/p/cen/wpaper/11-43.html.

———. 2011b. *LEHD Infrastructure Files in the Census RDC: Overview of S2004 Snapshot.* Working Papers 11-13. Center for Economic Studies, U.S. Census Bureau, April. http://ideas.repec.org/p/cen/wpaper/11-13.html.

National Science Foundation. 2011. “Data Management Plan.” Chap. II.C.2.j in *Grant Proposal Guide,* II–19. NSF, 11-1. National Science Foundation. Accessed April 16, 2014. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_index.jsp.

Review of Economics and Statistics. 2014. “Data Availability Policy.” Accessed April 16, 2014. http://www.mitpressjournals.org/page/sub/rest.

Stephens, Bryce. 2007. “Essays on firm compensation policy and confidentiality protection and imputation in the Quarterly Workforce Indicators.” Ph.D., University of Maryland, College Park.

Stevens, David W. 2007. *Employment that is not covered by state unemployment insurance laws.* Technical paper TP-2007-04. LEHD, U.S. Census Bureau.

U.S. Office of Personal Management. 2012. *The Guide to Data Standards - Part B: Payroll.* Technical report. Accessed at http://www.opm.gov/feddata/guidance.asp on May 8, 2012. OPM.

Vilhuber, Lars, and Kevin McKinney. 2014. *LEHD Infrastructure files in the Census RDC - Overview.* Working Papers 14-26. Center for Economic Studies, U.S. Census Bureau, June. https://ideas.repec.org/p/cen/wpaper/14-26.html.

# Appendix B.
# Errata

This section will contain a list of any errata found.

- 2018-09-24: file listings for QWI-SEINUNIT files (Section

## B.1 CHA:QWI

) erroneously mentioned `firmage` and `firmsize`. This has been corrected in the documentation.

```
$Id: overview_master.tex 11747 2014-06-20 14:48:21Z vilhuber $
```